Novelty Detection from an Ego-Centric Perspective

Omid Aghazadeh, Josephine Sullivan and Stefan Carlsson
Computer Vision and Active Perception Laboratory*
KTH, Sweden
{omida,sullivan,stefanc}@csc.kth.se

Abstract

This paper demonstrates a system for the automatic extraction of novelty in images captured from a small video camera attached to a subject's chest, replicating his visual perspective, while performing activities which are repeated daily. Novelty is detected when a (sub)sequence cannot be registered to previously stored sequences captured while performing the same daily activity. Sequence registration is performed by measuring appearance and geometric similarity of individual frames and exploiting the invariant temporal order of the activity. Experimental results demonstrate that this is a robust way to detect novelties induced by variations in the wearer's ego-motion such as stopping and talking to a person. This is an essentially new and generic way of automatically extracting information of interest to the camera wearer and can be used as input to a system for life logging or memory support.

1. Introduction

In this paper we address the problem of selecting and storing relevant parts of the visual input collected from a continuously worn camera capturing images at video rate. This problem is partly dictated by applications such as life logging [3, 9, 1] and memory support systems for the disabled [5]. Especially in the design of efficient memory support, there is a large potential advantage in the automatic selection of relevant moments of one's daily visual experience.

Memory selection depends on several factors relating to the complex state of the human observer and these are not primarily related to vision. Given just the visual input, however, we can ask ourselves which moments of the input we would like to capture and store and if there are any rules that can be formulated for this.

It is generally accepted that *novelty* is very central in deciding whether to remember something or not. It is a very natural criterion for selection both on pure data storage grounds as well as for the purely subjective reasons of later inspection of stored images. Heuristically novelty can be measured as the deviation from some standard background. The less variation there is in the background the easier it will be to detect novelty. One way to ensure that the background variation is limited is to choose a specific context within which novelty is selected.

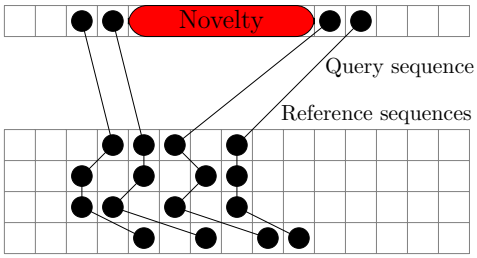


Figure 1: Novelty detection via sequence alignment.

Here we choose the simple context of the daily repeated activity of going to work. The collected video sequences from various days therefore contain image frames captured from approximately the same location. The influence of the day-to-day variation of these locations can be further reduced by aligning corresponding frames from different days using appearance and geometry information in the image frames. The content of a recorded sequence depends on two main factors: 1) the ego motion of the person wearing the camera and 2) the environment in which the sequence is captured. If there is a sufficient variation in one of these factors, this leads to the inability to register some or all of the sequence to previously stored sequences. This inability is taken as a measure of novelty. Ideally variations such as the person deviates from his/her daily path or stops to do some shopping or a street being shut off should be captured by our system.

This work extends previous studies based on wearable cameras in two main ways: 1) We use a very small (4cm high) camera that captures image at video rate for one hour and stores it on a memory stick. 2) Video is captured from

*This work was supported by The Swedish Foundation for Strategic Research in the project "Wearable Visual Information Systems".

daily repeated activities such as going to work and we develop algorithms for the automatic frame to frame registration of sequences recorded on different days. 3) We define novelty based on the absence of a good registration between a sequence and stored reference sequences.

The rest of paper is organized as follows. We begin by presenting in section 2 the details of our sequence alignment algorithm. This algorithm establishes frame to frame correspondences between two sequences. Section 3 then describes how the correspondences between sequences can be utilized to detect novelties. Afterwards, we present evaluation of the components of the proposed algorithm in section 4. Section 5 shows the results of the novelty detection algorithm and finally, section 6 concludes the paper.

2. Sequence alignment

Figure 2 displays 10 sequences from our dataset. Each row corresponds to one sequence and is of the subject walking from the a metro station to his work place. All our sequences are frames sampled from 25Hz videos at 1Hz. We wish to put the frames of one sequence s_1 in correspondence with another sequence s_2. As the sequences we capture have temporal continuity characteristics and repeated underlying structures, a natural way to establish correspondences is with Dynamic Time Warping (DTW). This algorithm requires a measure of similarity between each frame of s_1 and each frame of s_2 and the rest of this section is mainly devoted to how we compute this.

2.1. Appearance based cues

The most straightforward approach to define a measure of similarity between two sequences is to represent each frame with a fixed length vector and compare the representative vectors with a kernel such as polynomial or minimum intersection kernel. In order to represent frames with a fixed length vector, a common approach is to model the distribution of some local visual words[1] disregarding their spatial information.

Local features are fixed length description of some local interest regions localized in different areas of an input image. SIFT [7] and its variations are one of the most commonly used region descriptors. Various methods detect interest regions based on different criteria such as the determinant of the Hessian or the Harris tensor. A thorough study of region descriptors and interest region detectors is performed in [12]. Alternatively, it is possible to densely sample the SIFT features on multiple scales from a spatial grid over the image.

The local features are afterwards aggregated in a fixed length vector representing the entire image. The Bag of Features(BoF), inspired by text processing techniques, clusters features from many images to C clusters and models the frequency of assignments of the features in each image to one of cluster centers. This gives rise to a sparse C dimensional vector for each image regardless of the dimensionality of the features themselves. Recently, the Vector of Locally Aggregated Descriptors (VLAD) [6] was introduced that aggregates all the feature vectors assigned to the same cluster center to reach a vector of the same dimension as the visual words and performs the same for all cluster centers. This leads to a dense dC dimensional feature vector where d is the dimension of the local features.

We use the fast and efficient fixed length representation of the image to find the nearest neighbors of each frame of a query sequence in reference sequences. We will compare the performance of the VLAD and BoF aggregation methods on interest region based and dense sampling of SIFT features in our dataset in section 4.1.

2.2. Geometric similarity

The appearance features, described in the previous subsection, highlight pairs of frames which contain the same local structures. However, they do not guarantee that the matched local structures occur in a geometrically consistent way. The features can be considered as geometrically consistent if there is a global transformation or there are certain constraints are fulfilled between the matched features' locations encoding the relative position and orientation of the camera viewpoints. The tried and tested way to check this, especially when one may encounter large displacements and rotations between the views, is via epipolar geometry and estimation of the Fundamental matrix [8].

Thus we estimate the epipolar geometry between two views. Our measure of similarity is then defined as the percentage of inliers, with respect to the estimated fundamental matrix, in an initial set of putative matches. It should be noted we use this measure of similarity between two frames as an absolute score in $[0, 1]$, not a means for re-ranking [10], which is independent of the other images.

Estimating epipolar geometry robustly and efficiently

The images we capture are of dynamic environments and from a moving, twisting platform. Therefore we frequently have to match views with significant amounts of occlusion and significantly different viewpoints. We thus estimate the fundamental matrix from a sparse set of noisy correspondences and robust estimation via a RANSAC variant.

Unfortunately RANSAC based methods require an exponential number of trials in the minimum number of points required to fit the model and worse than exponential trials in the ratio of outliers to inliers. Given the large amount of data we have to process, a careful implementation w.r.t. the computational demands is required. Therefore we

[1] We use the terms visual words and features interchangeably in this article.

Figure 2: Each above row shows a highly temporally sub-sampled sequence from our dataset. Each sequence corresponds to a different day and captures what the subject experienced visually on his way to work.

- use Prosac [2] as it provides a significant speed up on RANSAC in the presence of a large number of outliers but where some inliers can be readily identified,
- reduce the minimum number of correspondences required to estimate the fundamental matrix from the standard 7 [4] to 5 by using the method suggested in [11] (though it does give up to 10 solutions),
- reduce the number of false correspondences in the initial putative set by choosing distinctive correspondences. As suggested in [7], we compute for each feature in one view the ratio of the Euclidean distance to its nearest neighbor and second nearest neighbor in the other view. These scores are sorted into ascending order and the first 250 features and its nearest neighbor match, w.r.t. this ordering, make up the putative set.

Another issue which has to be addressed is that the epipolar constraint is relatively weak (it maps a point in one view to a line in the other). To accurately judge the correctness of a hypothesized fundamental matrix in the presence of many incorrect correspondences additional constraints are needed. To this end we enforce that inliers must also be consistent with a homography mapping the local feature locations from one frame to the other. This homography consistency constraint is only weakly enforced and is achieved by using Prosac with a loose definition of inlier to robustly estimate a homography. Then only the matches which are consistent with this estimated homography are maintained and used for the fundamental matrix estimation. Algorithm 1 summarizes the complete implementation and the second row of figure 3 depicts the stages of the fundamental matrix estimation.

Algorithm 1 Computation of geometric similarity.

INPUT: Features $\mathcal{F}_1, \mathcal{F}_2$ extracted from images I_1, I_2
OUTPUT: Similarity measure $F_{GV}(I_1, I_2) \in [0, 1]$
$P \leftarrow N$ best putative matches between $\mathcal{F}_1$ and $\mathcal{F}_2$
$H_L \leftarrow$ PROSAC 4 points loose Homography(P)
$P_H \leftarrow$ inliers of P to H_L
$E \leftarrow$ PROSAC 5 point Essential Matrix(P_H)
$P_{HE} \leftarrow$ inliers of P_H to E
$F_{GV}(I_1, I_2) \leftarrow f_s(P_{HE}, P)$

The final geometric similarity measure

Once the fundamental matrix has been estimated and used to define a set of final point correspondences between the two views, we can calculate the geometric similarity score. In this work we define this as

$$f_s = \min\left(1, \alpha \max\left(0, \frac{|P_{HE}|}{|P|} - \beta\right)\right) \qquad (1)$$

where $|P|$ is the number of correspondences in the initial putative set and $|P_{HE}|$ is the number of final inliers found. The α and β are non-negative scalars which are learnt from training data. The role of β is to force the average matching score towards 0 for images which contain no overlap, while α scales the score with the aim that when images of the same scene are matched they achieve a score of around 1.

2.3. Dynamic time warping

Once one can measure similarity between two frames, using our geometric similarity measure, the temporal alignment of sequences is straightforward. There are just a cou-

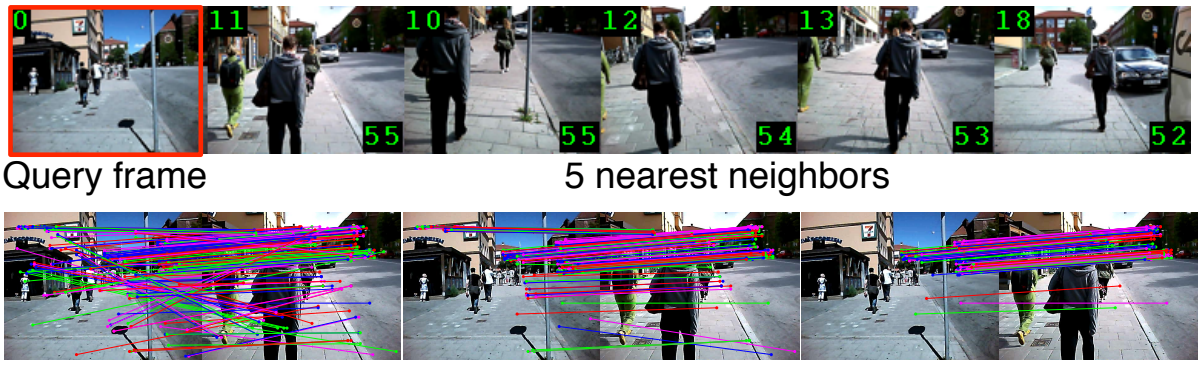


Figure 3: The top row shows the 5 nearest neighbors in a reference sequence to the query frame. While the bottom row shows the stages taken to establishing epipolar geometry between a query frame and a nearest neighbor. The initial correspondences are successively filtered by a robustly estimated homography and then the estimated epipolar geometry.

ple of steps involved. First the similarity matrix containing the similarity between any pairwise frames is formed and turned into a cost matrix by mapping the similarities to costs using a zero-mean Gaussian with standard deviation σ_c. Then temporal alignment is calculated via dynamic time warping on the cost matrix. Computing alignment in this fashion though straightforward is extremely slow as evaluating each entry in the cost matrix requires calculating the computationally expensive geometric similarity score. Clearly, it is not necessary to compute every entry, we just need to compute those which will have low costs.

These low cost entries can be easily identified, similar to [10], by utilizing the fast and efficient nearest neighbor search using the previously described appearance based fixed length representation, of the frames to find the k nearest neighbors in s_2 of each frame in s_1. Evaluation of the geometric similarity is then limited to k evaluations for each frame in s_1. As the same local features are used in the fixed length representation and in the geometric similarity evaluation, we expect the relevant low cost entries to be computed while ignoring the high cost entries. Figure 4 shows for one particular alignment example what proportion of geometric scores from the full matrix are actually computed and how the entries on the ground truth alignment path have been identified by the k nearest neighbor search.

The minimum cost path connecting the first and last entry of the cost matrix is denoted by a set of ordered pairs $\delta_{s_1,s_2} = \{(i_1, j_1), \ldots, (i_L, j_L)\}$ with $i_1 \leq i_2 \leq \cdots \leq i_L$ and similarly for the j's. We then define the *match cost* of a frame i in sequence s_1 to sequence s_2 as

$$\lambda(i, \delta_{s_1,s_2}) = \begin{cases} C_{i_k,j_k} & \text{if } \exists\,(i_k, j_k) \in \delta_{s_1,s_2} \text{ s.t. } i == i_k, \\ & i_k - i_{k-1} = 1 \textbf{ and } j_k - j_{k-1} = 1 \\ 1 & \text{otherwise} \end{cases} \quad (2)$$

where C_{i_k,j_k} is the value of the cost matrix at entry (i_k, j_k).

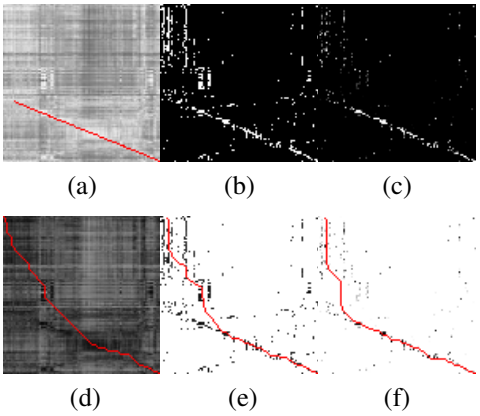


Figure 4: The similarity matrices calculated affect the ability to successfully align a sequence s_1 with another sequence s_2. **Top row:** (a) The full appearance similarity matrix and the ground truth registration between the two sequences is overlayed in red. (b) Sparse sampling of the appearance similarity matrix, using the 5 nearest neighbor per query frame (c) Sparse geometric similarity matrix, the geometric similarity is computed at non-zero entries of the b) matrix. **Bottom row:** The results of DTW applied to (d) dense appearance based cost matrix, (e) sparse appearance based cost matrix (f) sparse geometric similarity based cost matrix. Note how the final registration is closest to the ground truth.

Note the defined *match term* is unique for each frame due to the form of the path returned by dynamic time warping.

3. Novelty Detection

Once sequences can be aligned and correspondences can be established between their frames, then the quality of the alignments can be used for novelty detection. The crucial point is that novelties induce poor quality alignments. We therefore align a query sequence with the training sequences

and search for frames within the test sequence which do not have good correspondences in any or very few other sequences. Figure 1 illustrates such a situation.

Having aligned all sequences to a query sequence, for each frame of the sequence we compute the minimum match cost for each frame of the query sequence:

$$E(s_t^{(i)}) = \min_{s_r \in S} \ \lambda(i, \delta_{s_q, s_r}) \tag{3}$$

where S represents the set containing the reference sequences. If a frame has a good correspondence in at least one of the reference sequences, the entity $E(s_q^{(i)})$ will have a small value, otherwise it will have a bigger value close to 1. Therefore, we can directly threshold the *minimum match cost* to find novelties. A temporal smoothing of the minimum match cost E is applied prior to thresholding to reduce the effect to the multifarious sources of noise. We smooth the E's with a Gaussian mask with $\sigma_N = 2$ and then threshold them with $\theta_N = e^{-\frac{1}{2^3\sigma_c^2}}$ to detect novelties. This threshold is chosen as corresponds to a cost associated with a geometric similarity of 0.5.

4. Evaluation of the similarity matching

In this section we evaluate the quality of the performance of the constituent parts of the algorithm to compute the similarity between frames - the nearest neighbor search based on matching appearance and the geometric similarity scoring. It is crucial that these attain a certain level of performance to ensure that sequences can be registered in the presence of non-interesting variations. To help us do this we have manually annotated all sequences with a total of 9 different labels representing the location each frame of each sequences belongs to.

4.1. Nearest Neighbor search

The nearest neighbor search based on appearance features plays a critical role in creating the appropriate sparse cost matrix. Therefore we want to optimize its design and quantify its performance. There are numerous possible choices for the exact form of the features used and how they are compared as expounded in section 2.1. We limit, influenced by recent literature, our investigations to

- fixed length vector representations of the image with either BoF or VLAD descriptors built from SIFT features,
- the standard set of interest region detectors, see figure 5a, including a dense sampling[2].

Similarity between two images is then computed with the minimum intersection kernel for the BoF vectors and a polynomial kernel of degree one to compare VLAD vectors. When both representations are used, we use linear combination of the kernels with equal weights.

We then compare the label of a *query frame* with that of its K nearest neighbors and compute the proportion of the retrievals over the data set which return at least one correct label. Figure 5a shows the results of this experiment as the number of nearest neighbors returned and the image feature design varies. It can be observed that the dense sampling outperforms more specific interest region detection.

Guided by these results, we use the combination of the BoF and VLAD vectors with the color and gray variation in the final system. With this method, 88% of the time at least one of the 5 nearest neighbors to the query frame will correspond to a high similarity entry in the final cost matrix.

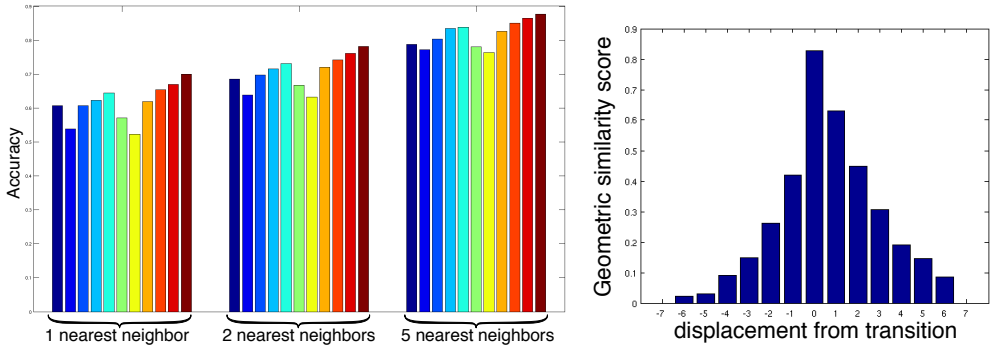
4.2. Geometric Similarity

There are many parameters that affect the performance of the geometric similarity function such as the number of fixed initial putative matches N, the thresholds θ_H and θ_E on the reprojection error for the estimated homography and essential matrix used to define inliers and the number of PROSAC iterations T_H and T_E used in estimating the homography and essential matrices. Although it is possible to find the configuration of the parameters by exhaustive search, such an approach would be extremely computationally expensive. Instead, we fixed the parameters and structure of F_{GV} empirically: we used $N = 250$, $\theta_H = 1$, $\theta_E = 0.01$, $T_H = 100$ and $T_E = 25$.

We evaluated the performance of the geometric similarity function using the dense sampling of the SIFT features and interest region detectors and found the dense sampling approach to perform better in terms of robustness and accuracy. This happens as 1) too many interest regions are found around the dynamic objects in the scene and these do not have a correspondence in the other frame and 2) too few interest regions are found in many regions which do not contain strong texture/gradients *e.g.* the the pavement in a relatively low resolution image. In these cases, it is no surprise that dense sampling approach can better capture information from the entire image.

The F_{GV} scores of a frame at a label transition matched to each frame in a local time window around a label transition to the same label as our target frame are computed and recorded. This process is repeated for all such transition frames and time windows. Figure 5b depicts the average result of this computation. On average F_{GV} maps the correct correspondence (the transition point)to a number close to 1 while its value drops monotonically relatively quickly with the displacement from the transition point. The appearance based fixed length representations would have a much slower drop and would not be able to precisely locate the label transitions as precisely or unambiguously.

[2]We use the implementation of dense SIFT features [13] with 4 scales and skip parameter of 6 pixels.



(a) (b)

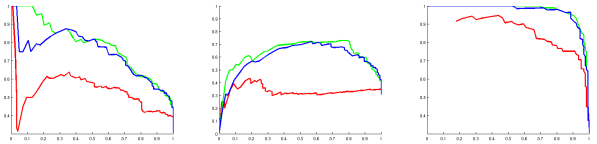
Figure 5: **(a)** The accuracy of image matching for differing interest region detectors and numbers of nearest neighbours. Methods (from left to right): VLAD+HessianAffine, VLAD+MSER, VLAD+HarrisAffine, VLAD+Dense(gray), VLAD+Dense(color), BoF+HessianAffine, BoF+MSER, BoF+HarrisAffine, BoF+Dense(gray), BoF+Dense(color), VLAD+BoF+Dense(gray+color). **(b)** The average of 100 F_{GV} values on local windows around the true correspondences.

5. Results

For the experiments in this paper, we used a data set of 31 sequences of the subject walking from metro station to work. In addition to the labelings mentioned earlier, we also manually defined temporal segments of the sequences in which something happened that either did not happen in the other sequences or it was infrequent *e.g.* subject meeting with a friend. The labelings resulted to 4 of the 31 sequences containing novel segments. Below, we present the results of the suggested algorithm trying to detect these 4 temporal segments.

Figure 7 depicts the intermediate and final results of novelty detection for a sequence containing novelty(the subject meets a friend). Due to limited space, we show the final picture containing 15 samples of the sequence(Figure 7a) vs 6 reference sequences. It can be observed that the method is able to detect both segments that were manually labeled as novel segments in addition to one false detection of a segment containing 4 frames(Figure 7d). The false detection is due to a very strong change in the lighting leading to a few overly bright frames; this inevitably leads to significant changes in local features which then prevents the algorithm to establish correct correspondences for those frames. Figure 8) depicts the results of novelty detection on the remaining 3 sequences that contain novel segments.

The accuracy of the novelty detection on 400 frames(4 sequences sub sampled to contain 100 frames) for which we had the ground truth manually labeled, are measured and depicted in figure 6. It can be observed that using dense the appearance costs leads to better accuracy compared to its sparse version. The figure also suggests that using the method with geometric costs outperforms the use of the appearance based costs with a strong margin. The high average precision of the results using geometric costs with as few as 6 reference sequences($AP \geq 0.96$)(green and blue curves in Figure 6c), suggests that the method is accurate and reliable for the purpose of novelty detection while being robust to various environmental changes such as view point and illuminations changes as well as occlusions.



(a) (b) (c)

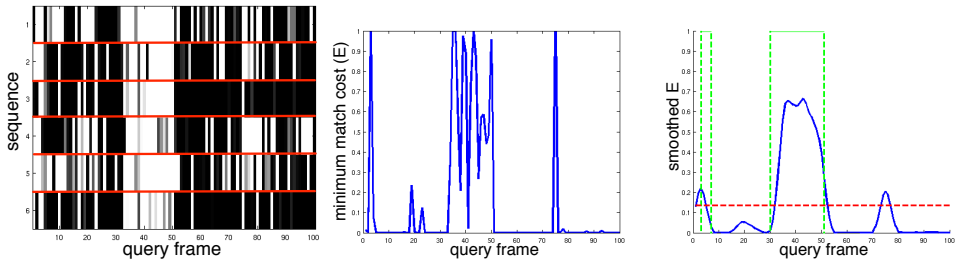
Figure 6: Precision-recall curves for novelty detection. Each figure uses a different cost matrix: (a) dense appearance , (b) sparse appearance, (c) sparse geometric. The red, green and blue curves show when 1, 6 and 10 reference sequences are used.

6. Conclusions and Future Work

We have demonstrated a system that is able to automatically extract novel events in the context of video captured from a camera continuously worn by a person who repeats a daily activity. The sequences manually annotated contain (subjectively) a total of four different novel events. All these novelties were automatically detected without any false positives. As far as we know this is the first systematic study of novelty detection of this kind where a repeated activity is used as background. These results indicate that potentially interesting applications of automatic memory selections should be possible especially in constrained environments like the kind considered here.

The frame-to-frame registration of the video captured from one day to another is possible, just using appearance

(a) The reference sequences are aligned with the query sequence and novelty is detected.



(b) (c) (d)

Figure 7: A detected novelty - *the subject meets a friend*. (a) The query frames without correspondences in the reference set, the black images below, are detected as novelties. Due to sub sampling only 3 of the 23 detected novelty frames are shown. (b) The *match cost* (λ) between each frame of the query sequence and the reference sequences it has been aligned to. Darker values correspond to lower costs. (c) The *minimum match cost* (E). (d) The smoothed *minimum match cost*. The red line shows the automatic threshold θ_N and the green curve the ground truth labeling of novelty. The large peak corresponds to the novelty displayed in figure (a).

and geometric cues, as we have constrained the variation in these sequences to those experienced by human wearer. This makes it possible to define a background relative to which novelty is measured.

In the future, we want to consider longer individual sequences captured over longer time periods. These will encompass many more activities in differing environments and will undoubtedly require a more complex description and representation of the captured background. Registration at a more abstract semantic level as opposed to the appearance/geometric level exploited in this paper will be needed. Novelty detection at a semantic level will allow disambiguation between false positives generated by changes in appearance and geometry induced by non-relevant variation of the environment or the ego-motion.

The central problem is the ability to measure similarity of recorded background with the actual captured video. In this sense the problem of novelty detection is intimately related to the general problem of similarity learning and the structuring of visual manifolds. We believe that the analysis of video captured from an ego-centric perspective can serve as an important test case for the study of these problems.

References

[1] U. Blanke and B. Schiele. Daily routine recognition through activity spotting. In *The 4th International Symposium on Location and Context Awareness*, 2009.

[2] O. Chum and J. Matas. Matching with prosac " progressive sample consensus. In *CVPR*, pages 220–226, Washington, DC, USA, 2005. IEEE Computer Society.

[3] A. Doherty and A. F. Smeaton. Automatically augmenting lifelog events using pervasively generated content from millions of people. *Sensors*, 10(3):1423–1446, 2010.

[4] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

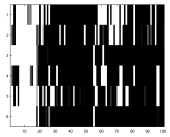
[5] S. Hodges, L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, and K. Wood. Sensecam: A retrospective memory aid. In *Proc. 8th International Conference on Ubicomp*, pages 177–193, 2006.

[6] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*, 2010.

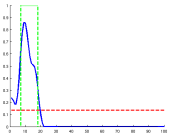
[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[8] F. Schaffalitzky and A. Zisserman. Automated scene matching in movies. In *In Proceedings of the Challenge of Image*

(a) *the subject runs into a friend*

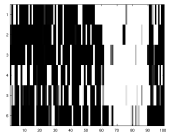


(b)

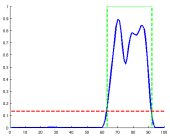


(c)

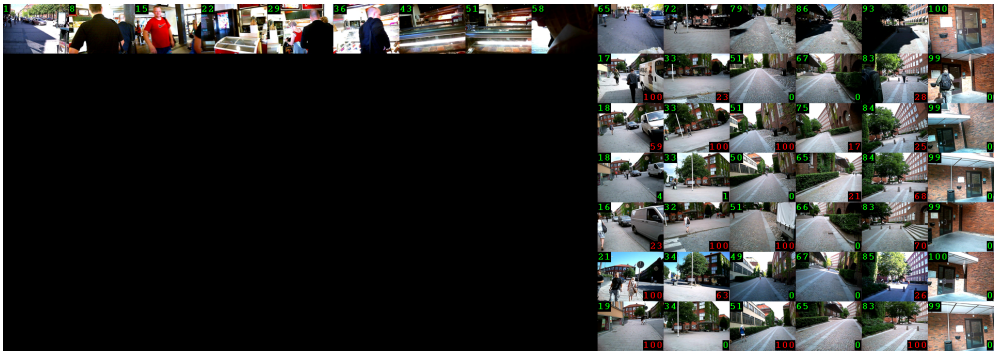
(d) *subject gives directions*



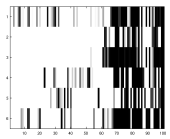
(e)



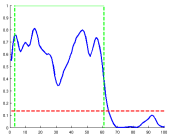
(f)



(g) *subject goes ice cream shopping*



(h)



(i)

Figure 8: Detected novelties in 3 sequences containing novelty and the corresponding *match costs* and smoothed *minimum match costs* on the right side.

and Video Retrieval, pages 186–197. Springer-Verlag, 2002.

[9] B. Schiele, N. Kern, and A. Schmidt. Recognizing context for annotating a live life recording. *Personal and Ubiquitous Computing*, 11(4):251–263, April 2007.

[10] J. Sivic and A. Zisserman. Video Google: Efficient visual search of videos. In *Toward Category-Level Object Recognition*, volume 4170 of *LNCS*, pages 127–144. Springer, 2006.

[11] H. Stewnius, C. Engels, and D. Nistr. Recent developments on direct relative orientation, 2006.

[12] T. Tuytelaars and K. Mikolajczyk. *Local Invariant Feature Detectors: A Survey*. 2008.

[13] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.