Context Change Detection for an Ultra-Low Power Low-Resolution Ego-Vision Imager

Francesco Paci[1], Lorenzo Baraldi[2], Giuseppe Serra[2],
Rita Cucchiara[2], Luca Benini[13]

[1] Univeristà di Bologna, Italy
{f.paci,l.benini}@unibo.it
[2] Università di Modena e Reggio Emilia, Italy
{lorenzo.baraldi,giuseppe.serra,rita.cucchiara}@unimore.it
[3] ETH Zürich, Switzerland
lbenini@iis.ee.ethz.ch

Abstract. With the increasing popularity of wearable cameras, such as GoPro or Narrative Clip, research on continuous activity monitoring from egocentric cameras has received a lot of attention. Research in hardware and software is devoted to find new efficient, stable and long-time running solutions; however, devices are too power-hungry for truly always-on operation, and are aggressively duty-cycled to achieve acceptable lifetimes. In this paper we present a wearable system for context change detection based on an egocentric camera with ultra-low power consumption that can collect data 24/7. Although the resolution of the captured images is low, experimental results in real scenarios demonstrate how our approach, based on Siamese Neural Networks, can achieve visual context awareness. In particular, we compare our solution with hand-crafted features and with state of art technique and propose a novel and challenging dataset composed of roughly 30000 low-resolution images.

Keywords: Egocentric Vision, ULP Camera, Low-Resolution, Deep Learning

1 Introduction and Related Works

Understanding everyday life activities is gaining more and more attention in the research community. This has triggered a number of interesting applications, ranging from health monitoring, memory rehabilitation, lifestyle analysis to security and entertainment [?,?,?,?]. These are mainly based on two sources of data: sensor and visual data. Sensor data, such as GPS, light, temperature and acceleration have been extensively used for activity monitoring [?,?,?]: among others, Kwapisz *et al.* [?] describe how a smartphone can be used to perform activity recognition simply by keeping it in the pocket. Guan *et al.* [?] present

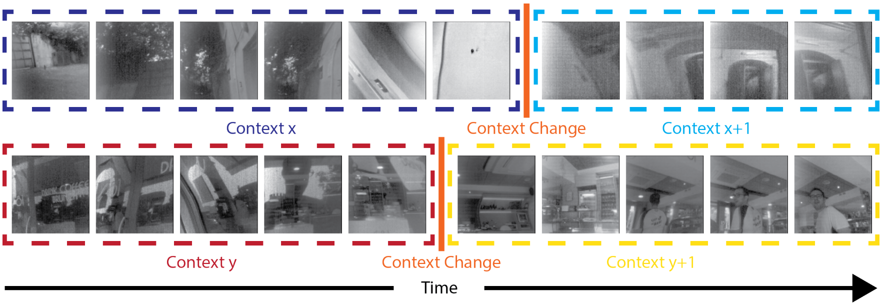


Fig. 1. We address the problem of recognizing context changes from low-Resolution images. Figure shows some images taken from the Stonyman Dataset.

a semi-supervised learning algorithm for action understanding based on 40 accelerometers strapped loosely to common trousers. Although sensor data can be easily collected for days, thanks to low energy consumption, its ability to recognize complex activities and the context around the user is low.

On the other hand, computer vision can indeed capture much richer contextual information which has been successfully used to recognize more complex activities [?,?,?]. Recently, several works that consider vision tasks from the egocentric perspective have been presented. Poleg *et al.* [?] propose a temporal segmentation that identifies 12 different activities (e.g. head motion, sitting, walking etc). Castro *et al.* [?] present an approach based on the combination of a Convolutional Neural Network and a Random Decision Forest; this approach is able to recognize images automatically in 19 activity classes. Ryoo *et al.* [?] suggest a new feature representation for egocentric vision which captures both the entire scene dynamics and the salient local motion observed in video. However, these approaches are designed to recognize a limited set of activities and can be useful for specific applications only.

To address this limitation, some unsupervised temporal segmentation and context change detection techniques have been presented, which are capable of splitting an egocentric video into meaningful segments. Lu *et al.* [?] present an approach that discovers the essential moments of a long egocentric video. First, they segment the original video into a series of subshots. Then they represent a short sequence in term of visual objects, that appear within it, using a bank of object detectors. Dimiccoli *et al.* [?] present an approach for context change detection, which combines low-level features and detection of semantic visual concepts (high-level semantic labels are extracted using Imagga's auto-tagging system[4]). By relying on these features, a graph-cut technique is used to integrate agglomerative clustering and an adaptive windowing algorithm [?].

All of these approaches exploit high quality videos and images taken by egocentric cameras that can be worn, like GoPro, Narrative Clip, Looxcie, Google Glass and Microsoft SenseCam. Although these cameras have become smaller

[4] https://imagga.com/solutions/auto-tagging.html

and cheaper, they are quite power-hungry. In fact, even if these devices take snapshots periodically, for example every 15 or 30 seconds, they have a short battery life ranging from one up to six hours. In addition, all presented solutions leverage imagers that, in the best case, consume several tens or hundreds of mW. These levels of power consumption are not affordable for continuous activity monitoring within a power envelope of a truly wearable system. Therefore, these solutions are not able to monitor human experience around the clock and their application in real contexts is limited in the analysis of short recording only.

We follow, therefore, another direction in contrast with the above mentioned. We explore how, even with very limited resolution, we can obtain context awareness and understand, at least, a change of context in our day-life. We present a context change detector for low-resolution images based on a wearable egocentric camera with ultra-low power consumption. An example of the task that we want to achieve is shown in Fig. 1. Low-resolution images can't "see" in the way we usually interpret, as good quality pictures, but can give visual context awareness, that can be exploited for context change detection. The system is able to collect data 24/7 laying the basis for the long-term analysis of egocentric vision activities. In this context, state of the art context change detection techniques, that are based on results of semantic classifiers, cannot be adopted. Therefore, we propose a novel approach that explores the use of Deep Convolutional Neural Networks on low level resolution images. Experimental results on a new challenging dataset demonstrate that the presented solution is able to detect context changes with good precision.

The paper is organized as follows. Section 2 gives an overview of the hardware system employed and presents the images and the pre-filtering stage, Section 3 describes in depth the Network architecture, Section 4 details the performance and accuracy of our solution, while Section 5 concludes the paper and gives some guidelines for future work.

2 Egocentric Vision Acquisition System

The egocentric vision acquisition system is based on a Texas Instrument Microntroller unit (MCU) and a low-power, low-resolution Stonyman Centeye imager. It is powered by a Li-Ion battery and embeds an energy harvester, that can supply the system while in operation or recharge the battery while the system is in standby. The main advantage of this platform is that it can continuously operate with a total power budget that is compatible with a small energy harvester or with 3,5 days of lifetime with a small (1Ah) battery. This platform is a development of Infinitime device [?], a wearable bracelet with human body harvesting. In Fig. 2 we show the core platform components and a picture of the real device.

The computational unit is an up-to 16 MHz Microcontroller by Texas Instruments, the MSP430FR5969 [?]. This MCU can run in several low power states, turning off unused memories and peripherals, or scaling down the operating fre-

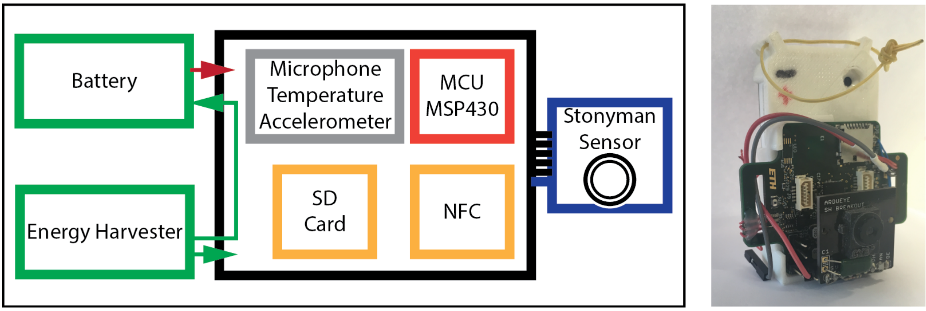


Fig. 2. Schema of the egocentric vision acquisition system.

quency. The sensors that this board features are the above mentioned imager, plus an analog microphone, a temperature sensor and an accelerometer.

2.1 Stonyman Imager

The embedded camera sensor, as already mentioned, is a Stonyman sensor by Centeye [?]. The first step to acquire images from this analog sensor is to sample them by an Analog to Digital Converter (ADC) and then store them in the system FRAM. Then the platform can store images in an SD card or send them through the NFC to a seconds device (e.g. a smartphone or a tablet).

The analog sensor can capture 112 × 112 pixel wide grayscale images at up to 2.5 fps, while storing it into SD card. The power consumption of the imager itself is orders of magnitude less than a digital CMOS sensors in the marketplace. In fact, we observed that the power consumption while reading an image is 3.9 mW, while storing an image into SD card takes about 121 mW. In terms of performance, acquiring an image and storing to SD card takes 400 ms. The sending procedure via NFC is less expensive in terms of power budget, as it costs 0.35 mW. In sleep mode the MCU consmes only 0,005 mW. So engaging a battery of 1000 mAH, at 1 fps and storing images in the SD Card the device can run for 3,5 days with a full recharge. Further experiments conducted by Spadaro *et al.* [?] shows that with a kinetic harvester during running activity the harvester can supply enough energy to collect 36 images per minute, while walking activity permits to take 6 images per minute using the NFC to send the image to second device. This ultra-low power consumption enables this device to be used as a perpetual visual aware sensor.

Images captured by the imager and converted by the ADC are rather noisy. A pre-filtering step is thus required to enhance the image quality. Next section discusses the image quality issues and the noise removal technique that we propose.

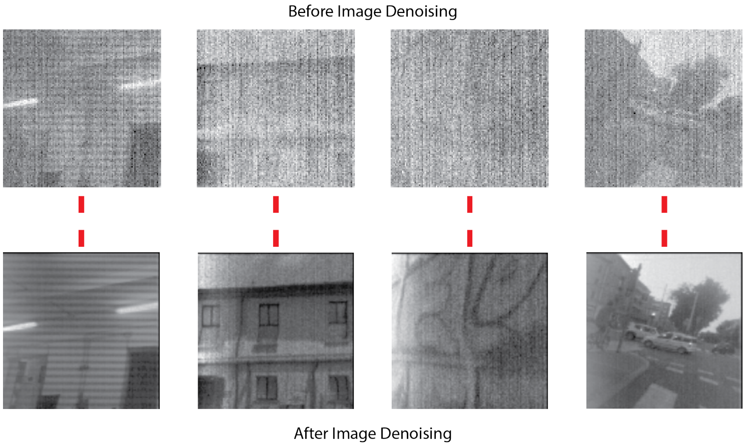


Fig. 3. Image Denoising results.

2.2 Images Pre-Processing

Images are sampled by a 12 bit ADC, so a normalization stage is needed to convert them in a 8-bit single channel format. In particular, images sampled from the Stonyman imager are mainly affected by static noise. Our noise removal system deals with it. Therefore, noise removal is carried out by subtracting a mask to the images, which is created by averaging several pictures framing a white background in an average light condition. In Figure 3 we show samples of images before and after denoising. After this stage, denoised images feed the temporal segmentation network described in next Section.

3 Temporal Segmentation Network

Learning to detect context changes can be addressed as a similarity learning task. In particular, we propose to learn a function $f(x, y)$ that compares an image x to another candidate image y of the same size and returns a high score if the two images capture the same context and a low score otherwise. The function f will be learned from a dataset of videos with labeled change points.

Given their widespread success in computer vision [?,?,?,?], we will use a deep ConvNet as the function f. The architecture of the network resembles that of a Siamese network [?], which is the most used model for addressing similarity learning with ConvNets. Siamese networks apply the same transformation ϕ to both inputs, and then combine their representations using a distance function. Therefore, function ϕ can be considered as an embedding, while the overall network can be seen as a learnable distance computation model.

To train the network, we employ a discriminative approach, by collecting positive and negative pairs. We define positive a pair of images which share the same temporal context, and negative a pair of images sampled from different contexts. At each iteration, we randomly sample a set of pairs $\mathcal{P}$, and minimize the following contrastive loss function:

$$L(\mathbf{w}) = \frac{1}{|\mathcal{P}|} \sum_{(x_i, y_i) \in \mathcal{P}} y_i f(x_i, y_i) + (1 - y_i) \max(0, 1 - f(x_i, y_i)) \tag{1}$$

where $y_i \in \{0, 1\}$ is the ground truth label of each pair. We choose to define the distance function f with respect to the embedding function ϕ through the cosine similarity:

$$f(x, y) = 1 - \frac{\phi(x) \cdot \phi(y)}{\|\phi(x)\| \cdot \|\phi(y)\|} \tag{2}$$

This choice, compared with more popular distance functions for Siamese networks, such as L_1 or L_2, presents a significant advantage. By computing the angle between $\phi(x)$ and $\phi(y)$, and neglecting their magnitudes, it does not force the network to bring its activations into a given numerical range, thus saving training time and avoiding poor local minima.

4 Results

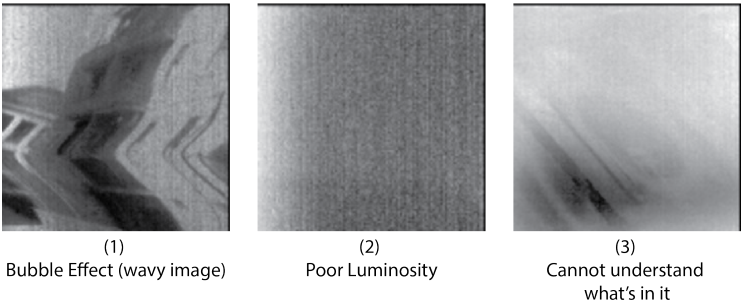
In this section we present the evaluation of our system in terms of accuracy in context change detection. The evaluation has been done by collecting a dataset of images that is described in the next section. In section 4.2 we describe the evaluation measures, while in Section 4.3 we present accuracy in comparison with two baselines and a state-of-art work.

4.1 Stonyman Dataset

To evaluate our results we collected a dataset of 29261 images named "Stonyman Dataset", from the name of the imager. All the images are collected at 1 fps and from a single subject under several days. We define context change any point of the sequence which delimits two temporal segments representing different

Table 1. Stonyman and Stonyman Quality Datasets: set names, number of images and number of context changes (CS).

Set Name	Stonyman D.	Stonyman Quality D.	# of CS
2016-04-06	2734	2143	7
2016-07-05	12104	9257	13
2016-07-06	6256	5566	11
2016-07-07 - 9.00	2056	1544	5
2016-07-07 - 12.00	4367	4043	6
2016-07-08	868	424	3
2016-07-09	876	435	3
Total	29261	23412	48



(1)
Bubble Effect (wavy image)

(2)
Poor Luminosity

(3)
Cannot understand
what's in it

Fig. 4. Three examples, matching the three criteria used to remove images from Stonyman Dataset to create Stonyman Quality Dataset.

environments (i.e. we considered as context change going in a shop, enter in the workplace, going off for a pause, catch the bus, etc).

In Table 1 we show the sets in which the dataset is divided and the number of images collected per day, while the third column shows the number of images of a subset of the dataset that we called "Stonyman Quality dataset". This is an improved version of the dataset obtained by pruning images with poor quality or that cannot be understood by a human expert. In particular, three criteria were considered:

1. Images with bubble effect (wavy images).
2. Images with poor luminosity or completely black.
3. Images where the subject that took the dataset cannot understand what's in it.

In Figure 4 an example of each of these defects is shown.

4.2 Evaluation metrics

For the evaluation of context scene detection, the classical precision-recall scheme has been often used, with the important variation of adding a temporal tolerance factor to detections and ground truth cuts. Therefore, a detection is considered as positive if its distance to nearest ground truth cut is below a certain threshold, otherwise it is considered as a false positive. False negatives are computed by counting ground truth cuts which are further than the same threshold to the nearest detected cut. Formally, given a threshold θ, a set of detected change points $D = \{t_0, t_1, ..., t_n\}$ and the set of ground truth cuts $C = \{t^g_0, t^g_1, ..., t^g_m\}$, true positives, false positives and false negatives are computed as follows:

$$TP = \sum_{i=0}^{n} \max_{j=0}^{m} 1(|t_i - t_j^g| \leq \theta) \quad FP = \sum_{i=0}^{n} 1 - \max_{j=0}^{m} 1(|t_i - t_j^g| \leq \theta) \tag{3}$$

$$FN = \sum_{i=0}^{m} 1 - \max_{j=0}^{n} 1(|t_j - t_i^g| \leq \theta)$$

where $1(\cdot)$ is an indication function that returns 1 when the given condition is true, and 0 otherwise. F-Score is then derived from Precision and Recall as usual.

Of course, the major drawback of this measure is the need to set an appropriate tolerance threshold. In our experiments, following previous works in the field [?], we set up a tolerance threshold of 5 frames, which given our frame rate correspond to 5 seconds.

The problem we address can be regarded as a temporal segmentation task, so appropriate measures can be taken from works that addressed temporal segmentation in other scenarios. One of them is surely scene detection, in which the objective is to temporally segment a broadcast video in semantically meaningful parts. In this settng, a measure based on intersection over union has been recently proposed [?]. Here, each temporal segment is represented as a closed interval, where the left bound of the interval is the starting frame, and the right bound is the ending frame of the sequence. The intersection over union of two segments a and b, $\mathrm{IoU}(a, b)$, is written as

$$\mathrm{IoU}(a, b) = \frac{a \cap b}{a \cup b} \tag{4}$$

A segmentation of a video can be seen as a set of non-overlapping sequences, whose union is the set of frames of the video. By exploiting this relation, [?] defines the intersection over union of two segmentations C and D as:

$$\overline{\mathrm{IoU}}(C, D) = \frac{1}{2}\left(\frac{1}{\#C}\sum_{a \in C} \max_{b \in D} \mathrm{IoU}(a, b) + \frac{1}{\#D}\sum_{b \in D} \max_{a \in C} \mathrm{IoU}(a, b)\right) \tag{5}$$

It is easy to see that Eq. 5 computes, for each ground-truth segment, the maximum intersection over union with the detected segments. Then, the same is done for detected segments against ground-truth ones, and the two quantities are averaged.

4.3 Experimental Results

To quantitatively evaluate the difficulty of dealing with low resolution images, we first present two baseline experiments. They both use Histogram of Oriented Gradients [?] (HOG) as descriptors, and hierarchical agglomerative clustering with euclidean distance to group images in contexts.

In the former baseline test (named CT1, Clustering Test 1), we fix the number of clusters to eight, which is the number of unique contexts that we have in our

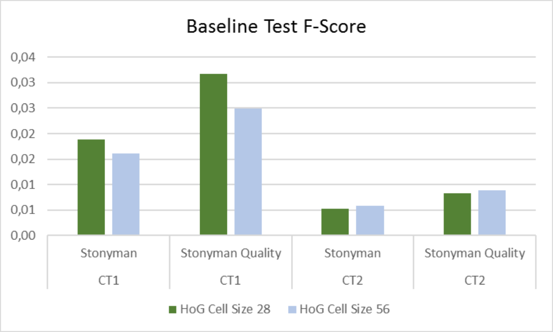


Fig. 5. CT1 and CT2 baselines in terms of F-Score

dataset: biking, car, home, office, walking, stairs, supermarket/shop, outdoor. The idea behind this experiment is to test the ability of a popular hand-crafted descriptor to distinguish between different contexts and places. HOG are extracted separately from each image and then descriptors are clustered in eight clusters.

In the latter test (named CT2, Clustering Test 2), instead, agglomerative clustering is applied with a different methodology, which resembles that of a Siamese network. Images are elaborated in subsequent couples from the beginning to end of the dataset. From each couple of images we extract HOG features, and compute the element-wise L_1 distance on feature vectors. We thus get a feature vector for each couple, having the same dimensionality of the HOG descriptor. The resulting features are then given as input to the agglomerative clustering, but instead of looking for eight clusters as the previous baseline test, we fix the number of clusters to two (similar and dissimilar pairs).

In Figure 5 and 6 we present the accuracy measured respectively with F-Score and IoU on CT1 and CT2. We tested two different settings for HOG features extraction. For both we used a window size of 112×112, block size of 56×56 and block stride of 28×28, and tested two different cell sizes: 28×28 and 56×56. We selected these two settings after conducting a grid search on a subset of the dataset, and picked the top two feature sizes in accuracy.

As it can be seen from the two charts, F-Score and IoU values are very low, thus revealing that hand-crafted features are not well suitable for low-resolution noisy images. The best accuracy in terms of F-Score is achieved with CT1, since the solution of clustering into eight classes is a more easy task, and we see a slight improvement with Stonyman Quality with respect to the entire dataset. In Figure 6 the same results are evaluated in terms of IoU. All settings results in similar values of IoU, and this is due to the completely different nature of the two performance measures.

Moving to the proposed approach, we employed the pre-trained 16 layers model from VGG [?] as the embedding function ϕ, since it is well known for its state-of-the-art performances on image classifications tasks, while still being

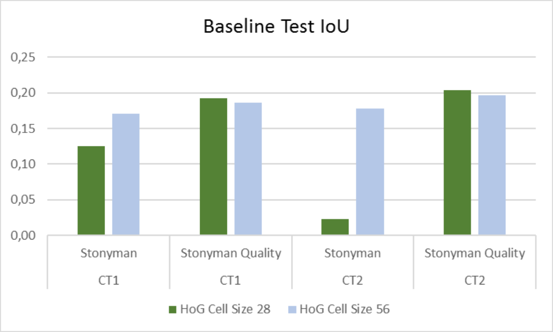


Fig. 6. CT1 and CT2 baselines in terms of IoU

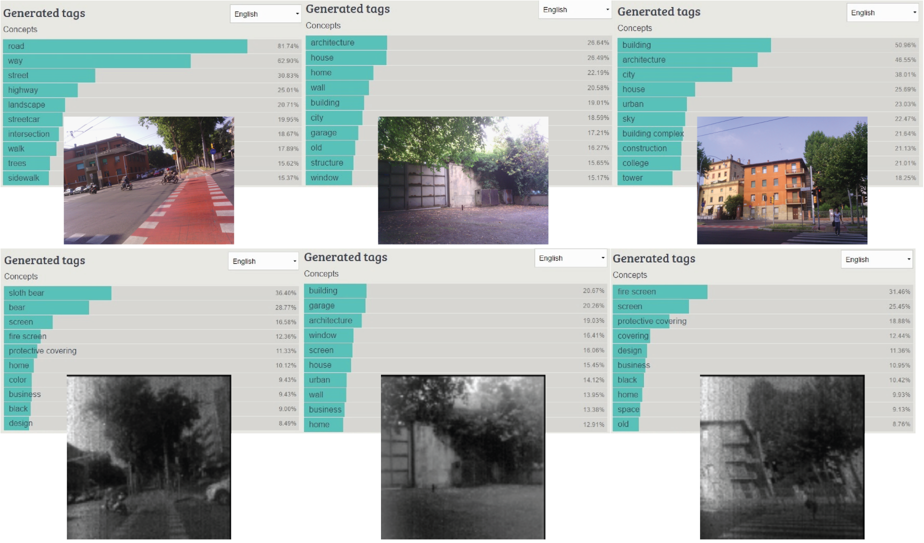
Table 2. F-Score and IoU results of our system on Stonyman and Stonyman Quality datasets

	Stonyman D.		Stonyman Quality D.	
	F-Score	IoU	F-Score	IoU
2016-04-06	0.571	0.655	0.667	0.608
2016-07-05	0.216	0.411	0.357	0.539
2016-07-06	0.105	0.590	0.286	0.375
2016-07-07 - 9.00	0.133	0.397	0.625	0.791
2016-07-07 - 12.00	0.217	0.387	0.500	0.712
2016-07-08	0.143	0.618	0.400	0.552
2016-07-09	0.193	0.346	0.267	0.520
Average	0.226	0.486	0.443	0.585

a simple and lightweight model for modern GPUs. The overall network is then trained end-to-end using Stochastic Gradient Descent with learning rate 0.001 and batches of 20 couples.

In Table 2 we present the results of our system on the Stonyman and Stonyman Quality datasets. The performances are reported in terms of F-Score and IoU for each set. Notice that Stoneyman Quality compared to Stonyman produce 0.2 improvement in F-Score and 0.1 improvement in IoU, we attribute this behavior mostly to wavy images that are removed in Stonyman Quality. These distort images produce an altered feature that make the problem more challenging.

Table 3 present a comparison of the two baselines (CT1 and CT2) and our system. We can observe that the techniques based on scene clustering achieve low performance. Whereas our system obtains promising results in both scenarios. We could not compare our solution with a state of the art Scene Clustering System called SR-Clustering [?], because, as mentioned before, a key element of their technique is the extensively usage of high-level semantic classifiers, which don't work with our low-resolution snapshots. This is clearly shown in Figure 7, in which we present some examples of predictions obtained on our low-resolution



road	81.74%
way	62.90%
street	30.83%
highway	25.01%
landscape	20.71%
streetcar	19.95%
intersection	18.67%
walk	17.89%
trees	15.62%
sidewalk	15.37%

architecture	26.64%
house	26.49%
home	22.19%
wall	20.58%
building	19.01%
city	18.59%
garage	17.21%
old	16.27%
structure	15.65%
window	15.17%

building	50.96%
architecture	46.55%
city	38.01%
house	25.69%
urban	23.03%
sky	22.47%
building complex	21.64%
construction	21.13%
college	21.01%
tower	18.25%

sloth bear	36.40%
bear	28.77%
screen	16.58%
fire screen	12.36%
protective covering	11.33%
home	10.12%
color	9.43%
business	9.43%
black	9.00%
design	8.49%

building	20.67%
garage	20.26%
architecture	19.03%
window	16.41%
screen	16.06%
house	15.45%
urban	14.12%
wall	13.95%
business	13.38%
home	12.91%

fire screen	31.46%
screen	25.45%
protective covering	18.88%
covering	12.44%
design	11.36%
business	10.95%
black	10.42%
home	9.93%
space	9.13%
old	8.76%

Fig. 7. Imagga predicted tags on the same images shot with Stonyman (Grayscale) and Narrative (Color)

images by the classifiers adopted in [?] and the corresponding Narrative Clip high quality image.

Table 3. Comparison results between the proposed solution and the two baselines (CT1 and CT2) on Stonyman and Stonyman Quality datasets.

	Stonyman D.		Stonyman Quality D.	
	F-Score	IoU	F-Score	IoU
CT1	0.019	0.170	0.032	0.192
CT2	0.006	0.179	0.009	0.204
Our System	**0.226**	**0.486**	**0.443**	**0.585**

Therefore even if our system cannot exploit an high-level semantic classifier, we tested it on the reference dataset of SR-Clustering to show that results on color high quality images are in-line with the Stonyman Quality low-resolution images.

The SR-Clustering work proposes a dataset called EDUB-Seg which is composed of two sets: EDUB-Seg Set 1 and EDUB-Seg Set 2. The only publicly available one is EDUB-Seg Set 1. This dataset is composed 4912 color images (512 × 512 pixels) collected by 4 subjects with a Narrative Clip camera [?] at 2 fpm. In Table 4 is shown our results on this dataset. We trained the network with the technique leave-one-out: for each subset the network is trained on all the other subsets. The results shows an improvement in F-Score compared to

Table 4. Performnace results of our system on EDUB-Seg

	F-Score	IoU
Subject1_1	0.563	0.494
Subject1_2	0.545	0.536
Subject2_1	0.448	0.466
Subject2_2	0.500	0.473
Subject3_1	0.500	0.418
Subject3_2	0.400	0.574
Subject4_1	0.476	0.546
Subject4_2	0.774	0.560
Average	0.521	0.510

Table 5. Accuracy of our system and SR-Clustering in EDUB-Seg Set 1 Dataset

	F-Score
SR-Clustering	0.69
Our System	0.521

Stoneyman Quality dataset, while on IoU there is a slight loss. This shows that in this dataset the low framerate is balanced by the quality of the images.

Lastly in Table 5 we report the results of SR-Clustering on EDUB-Seg set 1 and the average F-Score that we achieve with our system.

5 Conclusion and Further Work

In this paper we proposed a context change detection system. First, we presented an egocentric vision device with ultra-low power consumption that can capture images round the clock. Then, we suggested a similarity learning approach, based on Siamese ConvNets, that is able to deal with grayscale low-resolution snapshots. We finally run extensive experiments in real scenarios, showing the robustness and efficacy of the proposed method with respect to related approaches.

On future works we will explore an automatic technique for discarding images without a relevant semantic content. Moreover we will focus on embedding the network processing in an ultra-low power budget. The ultra-low power trend encourages deep-network based approaches [?] [?]. Experts in computer vision and also VLSI and computer architecture communities are focusing on these approaches with promising results in terms of energy efficiency.

Acknowledgements

This work was partially supported by the Swiss National Foundation under grant 162524 (MicroLearn: Micropower Deep Learning), the ERC MultiTherman project (ERC-AdG-291125) and the Vision for Augmented Experiences through the Fondazione CRMO Project.

References