Time-Sensitive Egocentric Image Retrieval for Finding Objects in Lifelogs

Author:

Cristian Reyes

Advisors:

Eva Mohedano

Kevin McGuinness

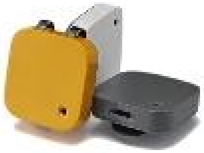
Xavier Giró

14th July 2016

Outline

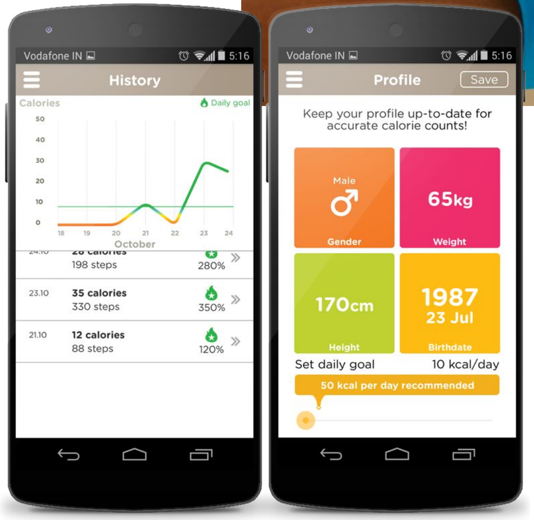
1. **Motivation**
2. Methodology
3. Experiments
4. Results & Conclusions

Lifelogging

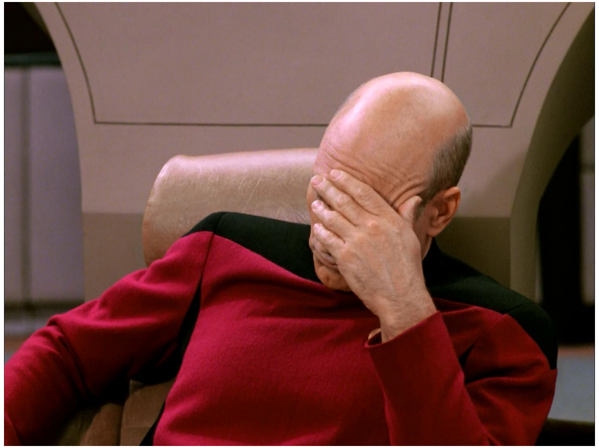


Extract value from this new data


Running
4:46
45:00
MOST: 60 MIN
Start


Vodafone IN
5:16
History
Calories
Daily goal
October
198 steps
280%
23.10
35 calories
330 steps
350%
21.10
12 calories
88 steps
120%
Vodafone IN
5:16
Profile
Save
Keep your profile up-to-date for accurate calorie counts!
Male
Gender
65kg
Weight
170cm
Height
1987
23 Jul
Birthdate
Set daily goal
10 kcal/day
50 kcal per day recommended

Motivation

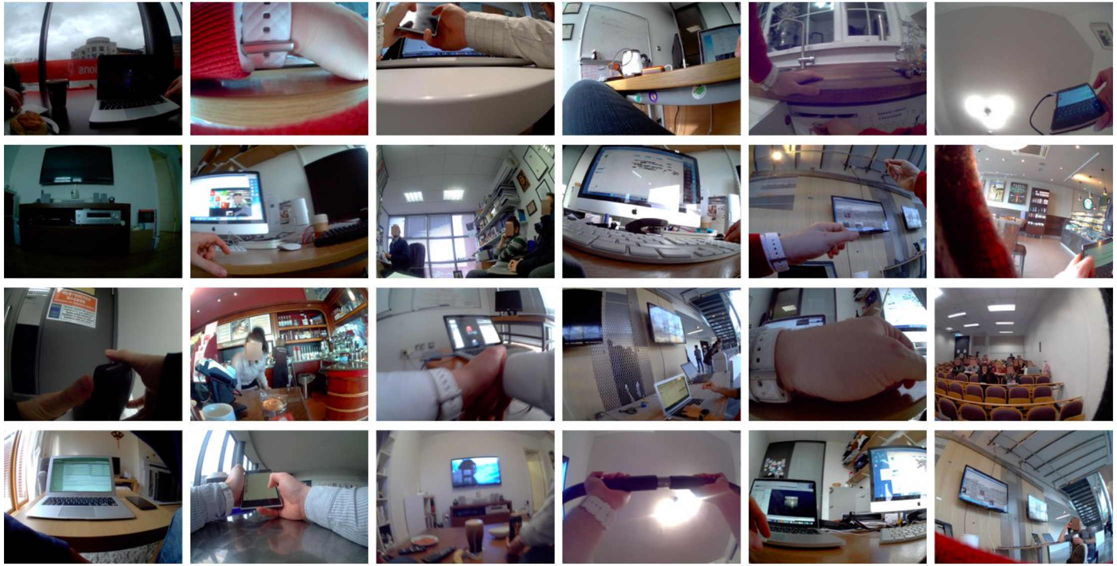


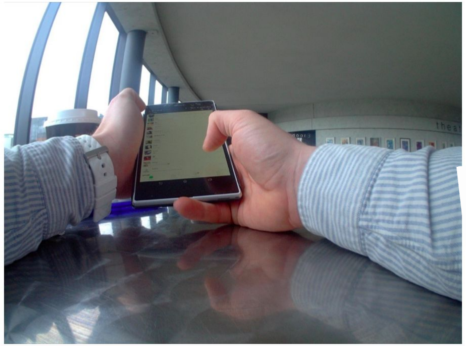
CAN'T FIND MY PHONE

Egocentric cameras may help

Review

Hundreds of images!



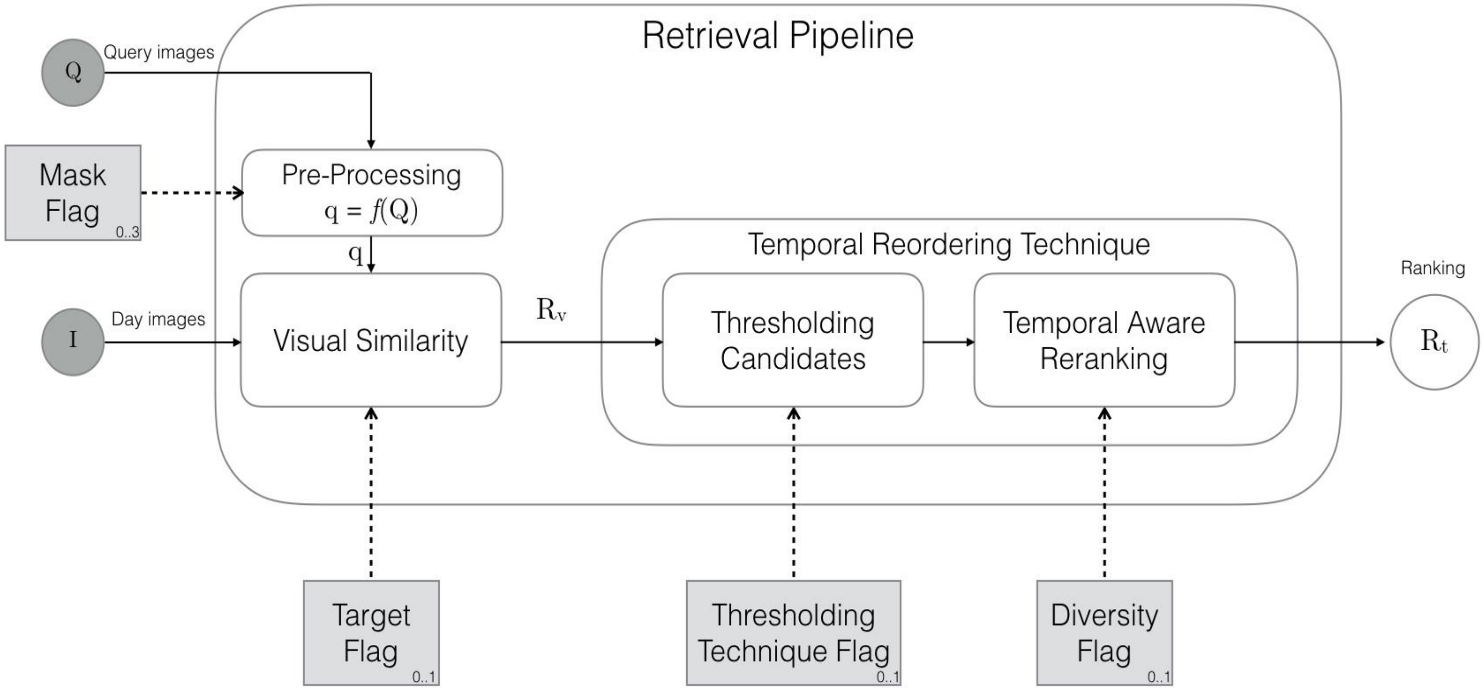


Last time seen: At the CAFE

Task

Goal: Retrieve a **useful image** to find the object.

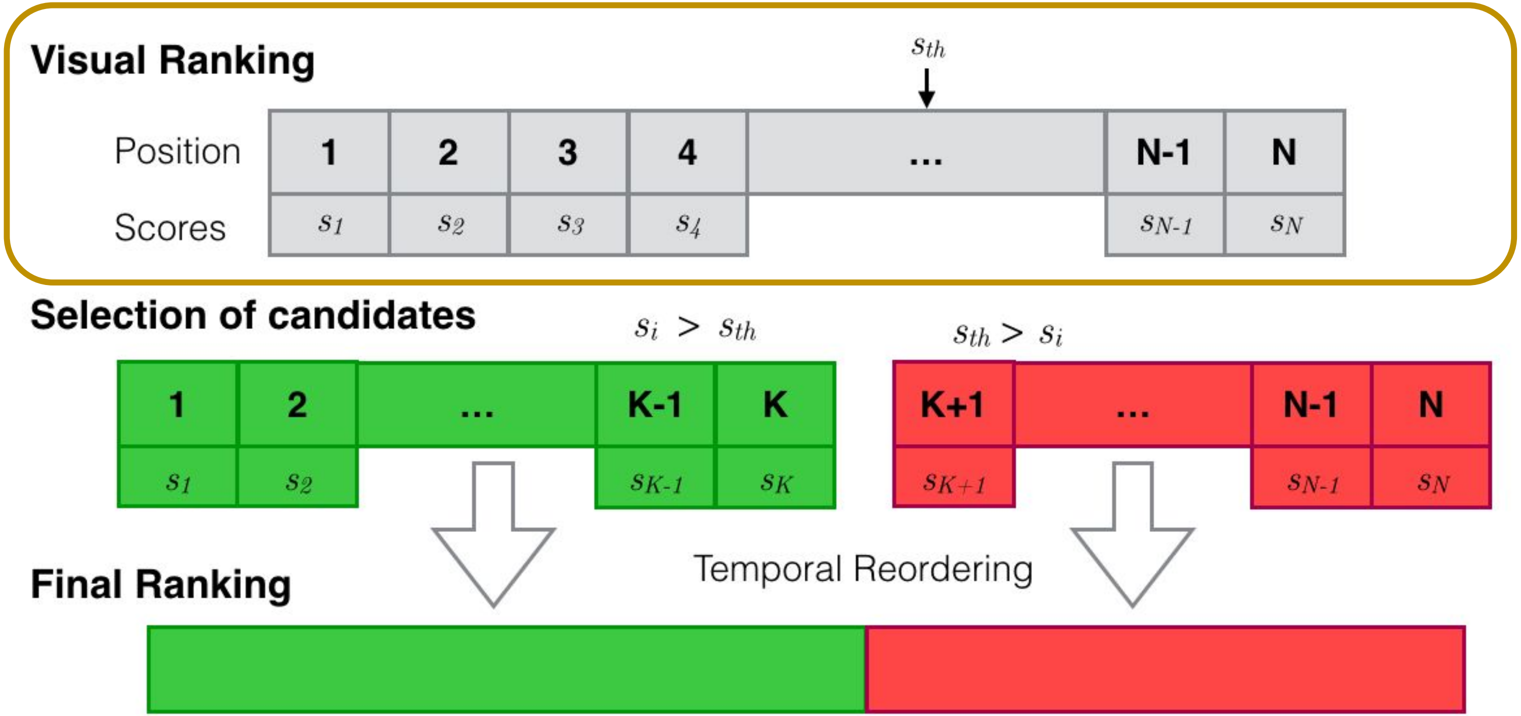
How: Exploiting **visual** and **temporal** information.



Outline

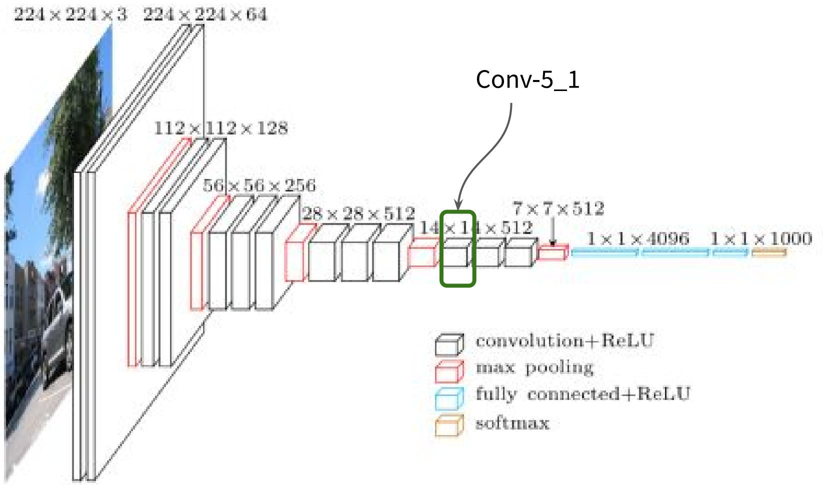
1. **Motivation**
2. **Methodology**
3. Experiments
4. Results & Conclusions

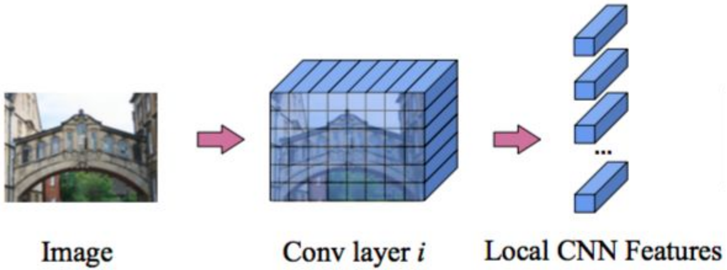
System Overview



Visual Ranking - Descriptors

Convolutional Neural Networks



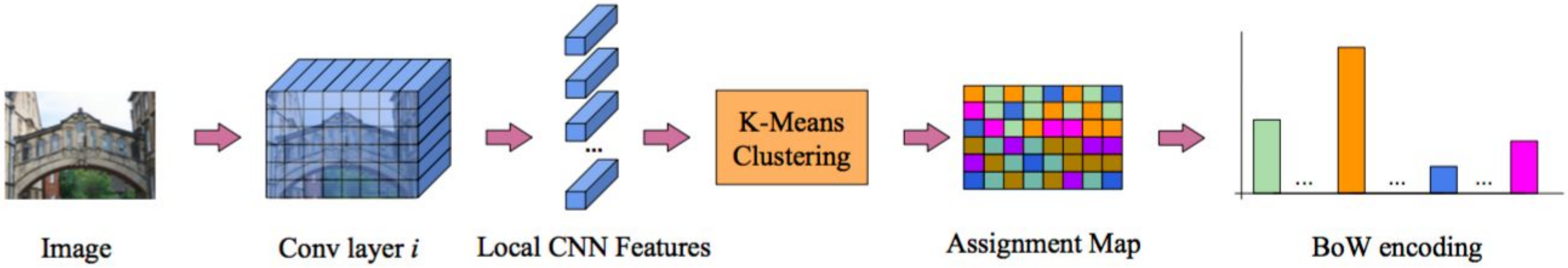


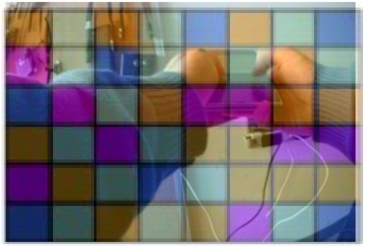
CNN design (*vgg16*) used for feature extraction.

Eva Mohedano, Amaia Salvador, Kevin McGuinness, Ferran Marques, Noel E. O'Connor, and Xavier Giro-i Nieto.
Bags of local convolutional features for scalable instance search. In Proceedings of the ACM International Conference on Multimedia. ACM, 2016.

Visual Ranking - Descriptors

Bag of Words

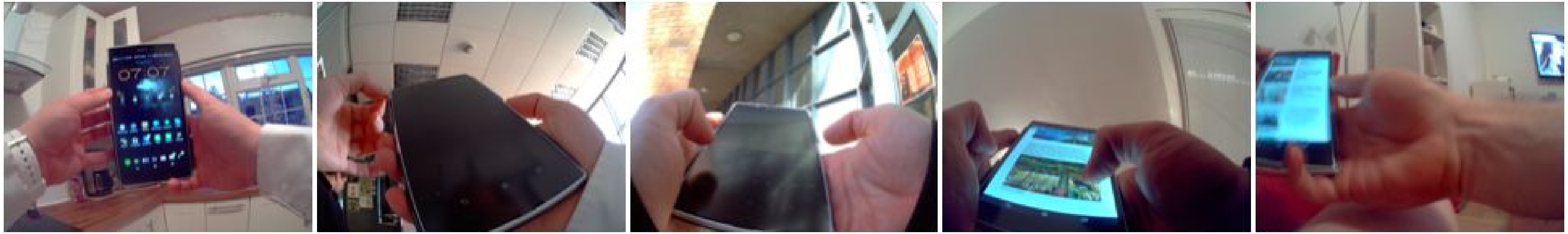




Eva Mohedano, Amaia Salvador, Kevin McGuinness, Ferran Marques, Noel E. O'Connor, and Xavier Giro-i Nieto.
Bags of local convolutional features for scalable instance search. In Proceedings of the ACM International Conference on Multimedia. ACM, 2016.

Visual Ranking - Queries

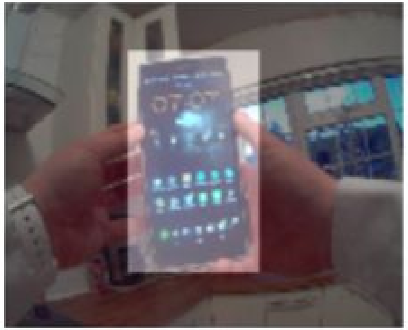
5 visual **examples**



Visual Ranking - Queries

3 masking **strategies**

Full Image
(FI)



Hard Bounding Box
(HBB)

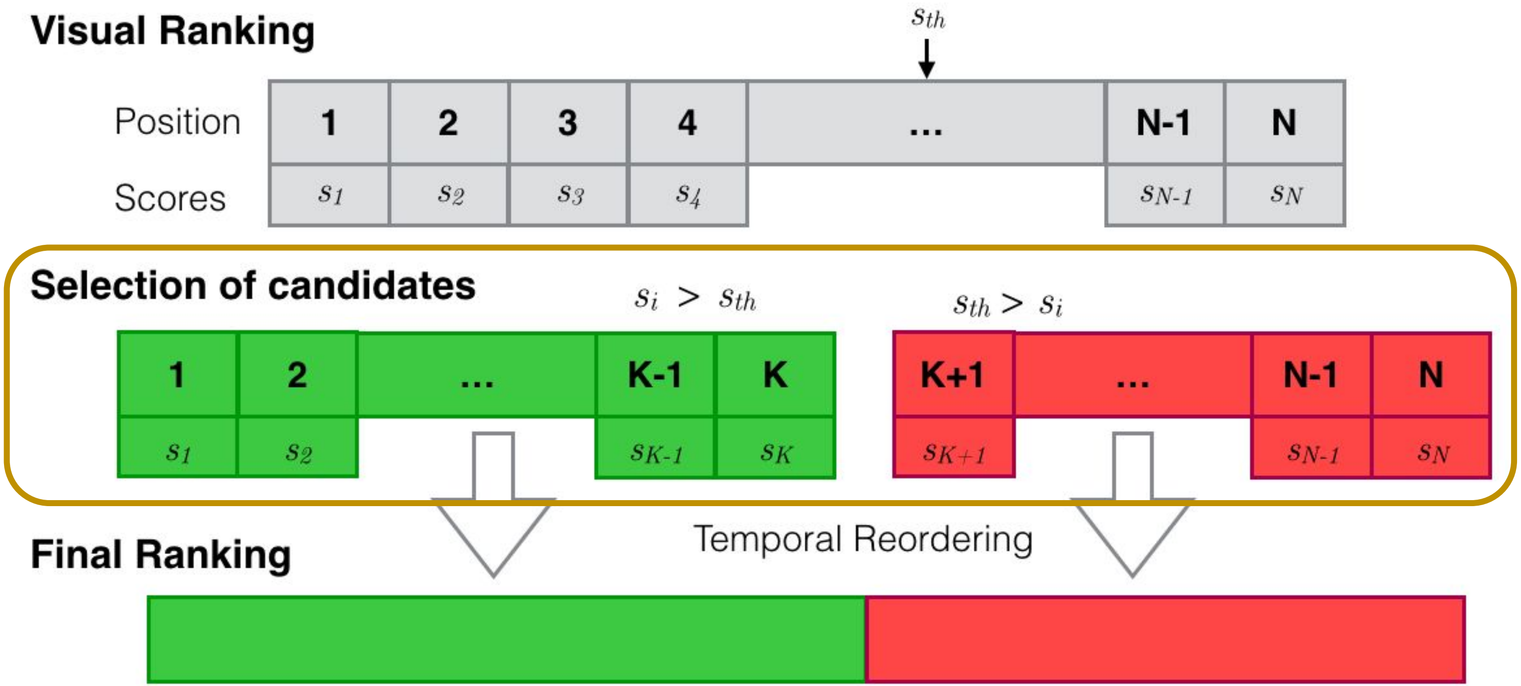
Soft Bounding Box
(SBB)

Visual Ranking - Target

3 masking **strategies**

Junting Pan, Kevin McGuinness, Elisa Sayrol, Noel O'Connor, and Xavier Giro-i Nieto.
Shallow and deep convolutional networks for saliency prediction. CVPR 2016.

System Overview



Candidate Selection

2 thresholding **strategies**

Absolute

$$C = \{p \in I : \nu_p > \nu_{th}\}$$

Threshold on Visual Similarity Scores (TVSS)

Adaptive

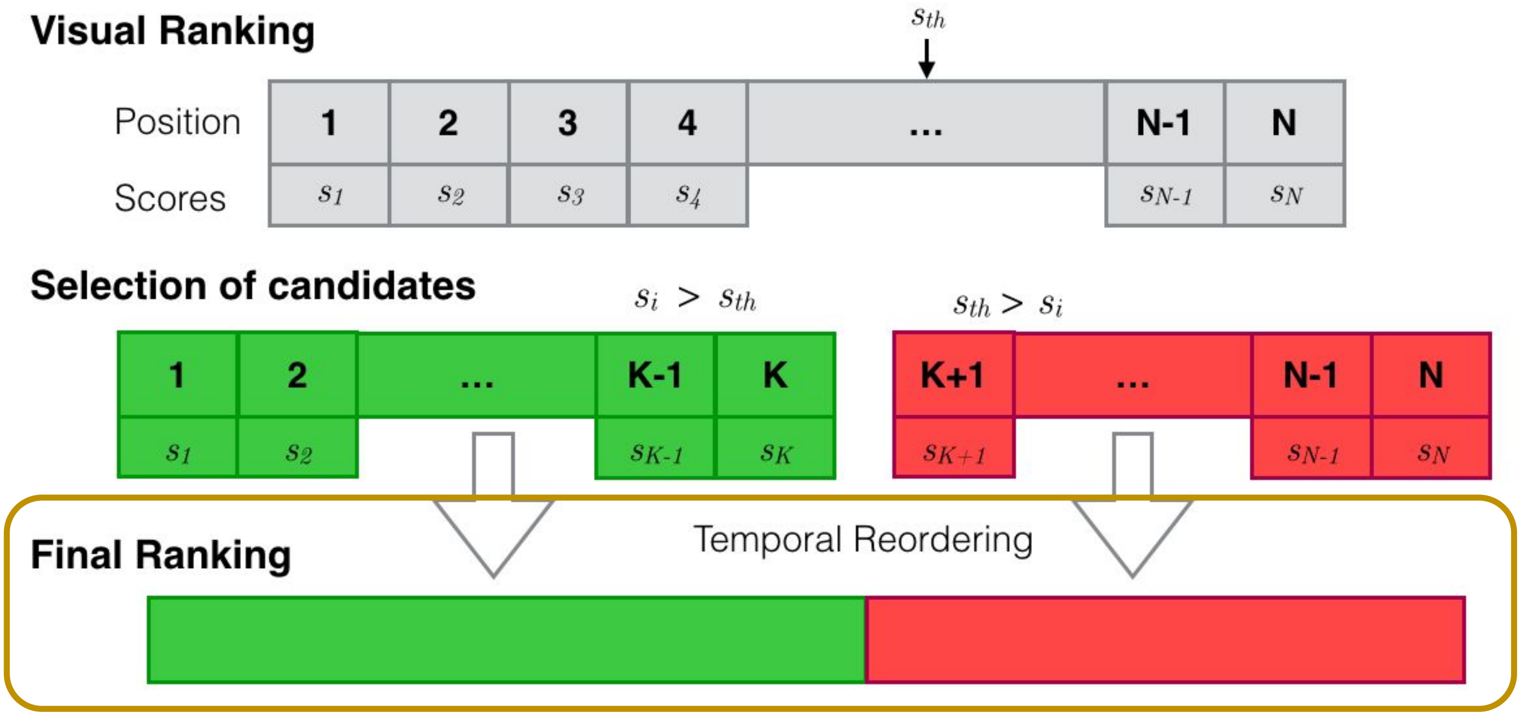
$$C = \left\{i \in I : \frac{\nu_i}{\nu_1} > \rho_{th} \frac{\nu_2}{\nu_1}\right\}$$

Nearest Neighbor Distance Ratio (NNDR)

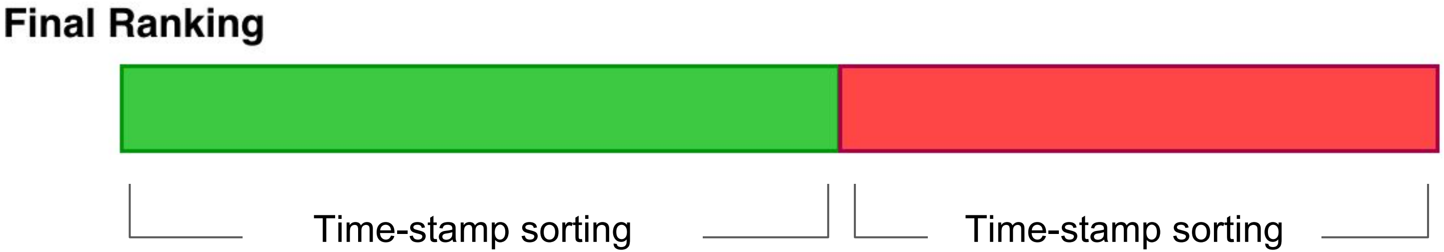
Parameters **LEARNT**

D. G. Loewe.
Distinctive image features from scale-invariant keypoints, 2004.

System Overview



Temporal aware reranking

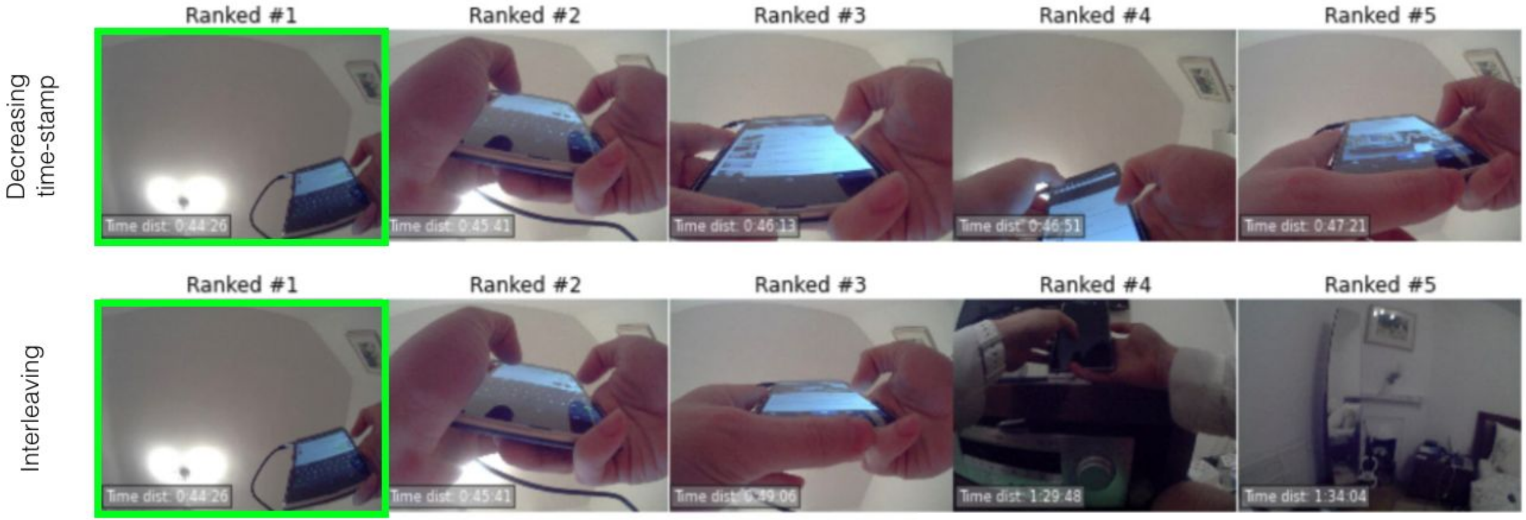


Redundancy


Decreasing time-stamp
Ranked #1
Ranked #2
Ranked #3
Ranked #4
Ranked #5
Time dist: 0:44:26
Time dist: 0:45:41
Time dist: 0:46:13
Time dist: 0:46:51
Time dist: 0:47:21

Temporal Diversity

Effect of Diversity



New locations introduced

Outline

1. **Motivation**
2. **Methodology**
3. **Experiments**
4. Results & Conclusions

Compare
different configurations

Evaluate
quantitatively

1. Dataset of images
2. Queries
3. Annotations
4. Metric

Dataset

EDUB[1]

- 4912 images
- 4 users
- 2 days/user
- Narrative Clip 1

NTCIR-Lifelog[2]

- 88185 images ✓
- 3 users
- 30 days/user ✓
- Autographer ✓

⇩

General Behavior
Wide Angle Lens

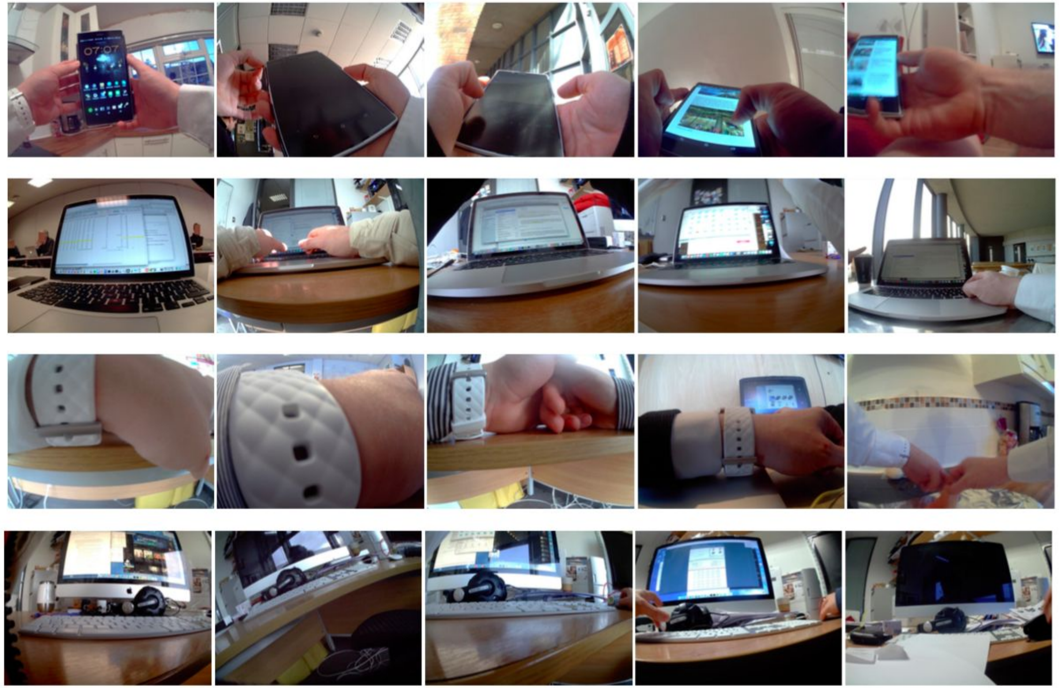
1. Marc Bolaños and Petia Radeva.
Ego-object discovery.

2. C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, and R. Albatal.
NTCIR Lifelog: The first test collection for lifelog research.

EDUB

NTCIR-Lifelog

Dataset - Query Definition



Mobile phone

Laptop

Watch

Headphones

Annotation Strategy

1. ~~Annotate the **3 last** occurrences~~ → The system is expected to find all 3

2. ~~Annotate **only** the last occurrence~~ → Neighbor images may also be helpful

3. Annotate **all the scene** of the last occurrence ✓

Metric

Mean Average Precision

$$\mathrm{MAP} = \frac{1}{|Q|} \sum_{q \in Q} \mathrm{AP}(q)$$

$$\mathrm{AP} = \frac{1}{|R|} \sum_{k=1}^{N} \mathrm{P}(k) \cdot \mathbb{1}_R(k)$$

ALL RELEVANT IMAGES

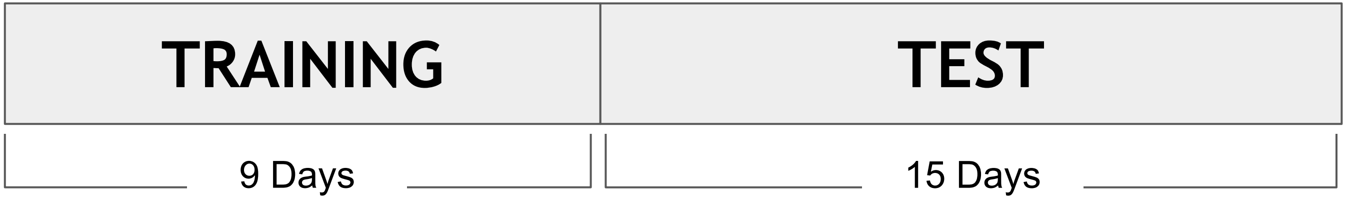
Mean Reciprocal Rank

$$\mathrm{MRR_d} = \frac{1}{|Q_d|} \sum_{q \in Q_d} \frac{1}{q^*}$$

$$\mathrm{A\text{-}MRR} = \frac{1}{|D|} \sum_{d \in D} \mathrm{MRR_d}$$

THE BEST RANKED RELEVANT IMAGE

Training



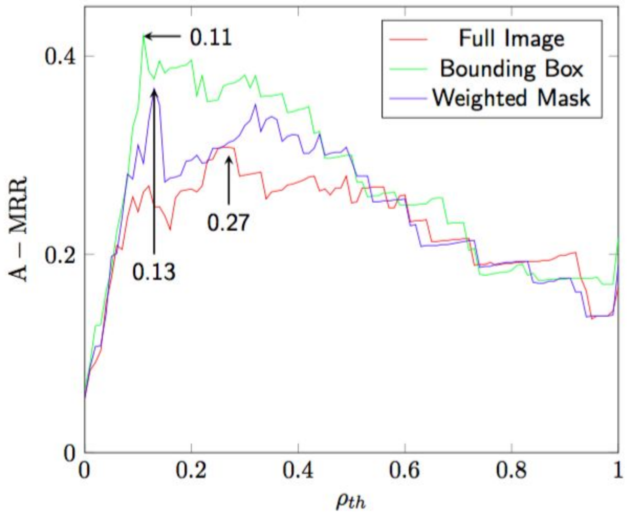
Training - Codebook

~ 13,500
images

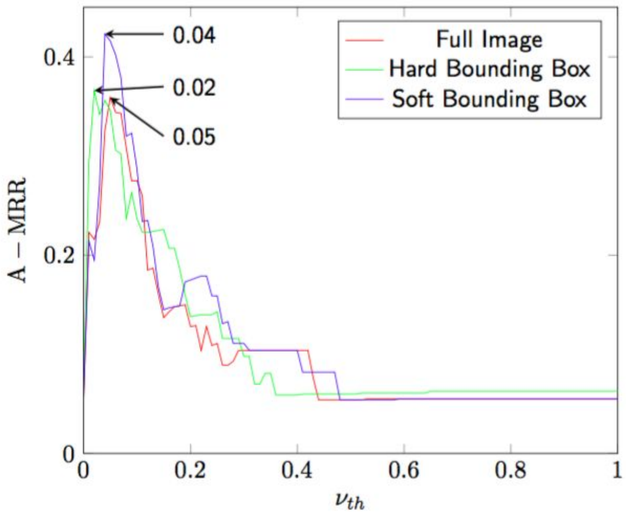
~18,000,000
feature vectors

25,000
clusters

Training - Thresholds



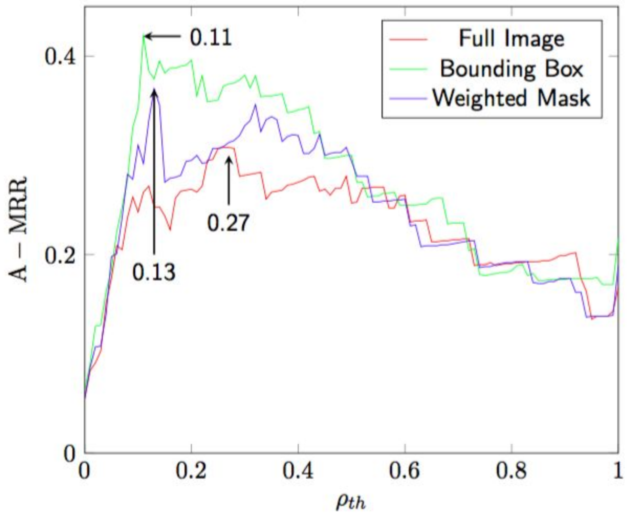
(a) NNDR, No Diveristy



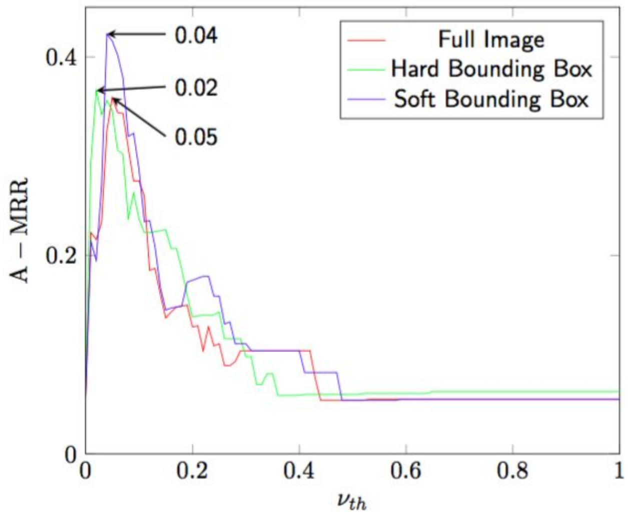
(b) TVSS, No Diveristy

Training - Thresholds

SAME OPTIMAL THRESHOLDS



(a) NNDR, No Diveristy



(b) TVSS, No Diveristy

TIME-STAMP SORTING
INTERLEAVING

FULL IMAGE
CENTER BIAS
SALIENCY MASK

Parameters summary

Flag	Possible Approaches
Query mask flag $f(Q)$	Full Image (FI) Hard Bounding Box (HBB) Soft Bounding Box (SBB)
Target processing flag $g(i)$	Full Image (FI) Center Bias (CB) Saliency Maps (SM)
Thresholding flag	Nearest Neighbor Distance Ratio (NNDR) Threshold on Visual Similarity Scores (TVSS)
Temporal reordering flag	Time-stamp Sorting Interleaving

Outline

1. Motivation
2. Methodology
3. Experiments
4. **Results & Conclusions**

Discussion & Conclusions

QUERY APPROACH

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I [5]	TVSS+I
FI		0,157	0,216	0,213	0,231	0,223
HBB	0,051	0,139	0,212	0,180	0,216	0,184
SBB		0,163	0,171	0,257	0,169	**0,269**

Table 4.2: A - MRR using Full Image for g.

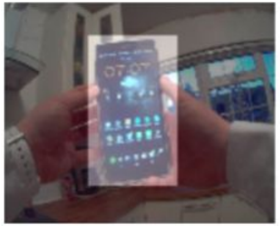
$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,156	0,191	0,205	0,206	0,215
HBB	0,051	0,130	0,212	0,170	0,216	0,174
SBB		0,162	0,160	0,240	0,161	**0,258**

Table 4.3: A - MRR using Center Bias for g.

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,150	0,240	0,274	0,249	**0,283**
HBB	0,051	0,173	0,200	0,136	0,206	0,147
SBB		0,178	0,168	0,242	0,174	0,257

Table 4.4: A - MRR using Saliency Maps for g.

FI



HBB

SBB

Discussion & Conclusions

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I [5]	TVSS+I
FI		0,157	0,216	0,213	0,231	0,223
HBB	0,051	0,139	0,212	0,180	0,216	0,184
SBB		0,163	0,171	0,257	0,169	**0,269**

Table 4.2: A - MRR using Full Image for g.

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,156	0,191	0,205	0,206	0,215
HBB	0,051	0,130	0,212	0,170	0,216	0,174
SBB		0,162	0,160	0,240	0,161	**0,258**

Table 4.3: A - MRR using Center Bias for g.

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,150	0,240	0,274	0,249	**0,283**
HBB	0,051	0,173	0,200	0,136	0,206	0,147
SBB		0,178	0,168	0,242	0,174	0,257

Table 4.4: A - MRR using Saliency Maps for g.

TARGET APPROACH

Full Image

Center Bias

Saliency Mask

Discussion & Conclusions

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I [5]	TVSS+I
FI		0,157	0,216	0,213	0,231	0,223
HBB	0,051	0,139	0,212	0,180	0,216	0,184
SBB		0,163	0,171	0,257	0,169	**0,269**

Table 4.2: A - MRR using Full Image for g.

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,156	0,191	0,205	0,206	0,215
HBB	0,051	0,130	0,212	0,170	0,216	0,174
SBB		0,162	0,160	0,240	0,161	**0,258**

Table 4.3: A - MRR using Center Bias for g.

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,150	0,240	0,274	0,249	**0,283**
HBB	0,051	0,173	0,200	0,136	0,206	0,147
SBB		0,178	0,168	0,242	0,174	0,257

Table 4.4: A - MRR using Saliency Maps for g.

Adaptive: NNDR

Absolute: TVSS

+

Time-stamp reordering

Discussion & Conclusions

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I [5]	TVSS+I
FI		0,157	0,216	0,213	0,231	0,223
HBB	0,051	0,139	0,212	0,180	0,216	0,184
SBB		0,163	0,171	0,257	0,169	**0,269**

Table 4.2: A-MRR using Full Image for g.

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,156	0,191	0,205	0,206	0,215
HBB	0,051	0,130	0,212	0,170	0,216	0,174
SBB		0,162	0,160	0,240	0,161	**0,258**

Table 4.3: A-MRR using Center Bias for g.

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,150	0,240	0,274	0,249	**0,283**
HBB	0,051	0,173	0,200	0,136	0,206	0,147
SBB		0,178	0,168	0,242	0,174	0,257

Table 4.4: A-MRR using Saliency Maps for g.

Adaptive: NNDR

Absolute: TVSS

+

Interleaving

Discussion & Conclusions

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I [5]	TVSS+I
FI		0,157	0,216	0,213	0,231	0,223
HBB	0,051	0,139	0,212	0,180	0,216	0,184
SBB		0,163	0,171	0,257	0,169	**0,269**

Table 4.2: A - MRR using Full Image for g.

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,156	0,191	0,205	0,206	0,215
HBB	0,051	0,130	0,212	0,170	0,216	0,174
SBB		0,162	0,160	0,240	0,161	**0,258**

Table 4.3: A - MRR using Center Bias for g.

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,150	0,240	0,274	0,249	**0,283**
HBB	0,051	0,173	0,200	0,136	0,206	0,147
SBB		0,178	0,168	0,242	0,174	0,257

Table 4.4: A - MRR using Saliency Maps for g.

Discussion & Conclusions

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I [5]	TVSS+I
FI		0,157	0,216	0,213	0,231	0,223
HBB	0,051	0,139	0,212	0,180	0,216	0,184
SBB		0,163	0,171	0,257	0,169	**0,269**

Table 4.2: A - MRR using Full Image for g.

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,156	0,191	0,205	0,206	0,215
HBB	0,051	0,130	0,212	0,170	0,216	0,174
SBB		0,162	0,160	0,240	0,161	**0,258**

Table 4.3: A - MRR using Center Bias for g.

The system **helps** the user

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,150	0,240	0,274	0,249	**0,283**
HBB	0,051	0,173	0,200	0,136	0,206	0,147
SBB		0,178	0,168	0,242	0,174	0,257

Table 4.4: A - MRR using Saliency Maps for g.

Discussion & Conclusions

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I [5]	TVSS+I
FI		0,157	0,216	0,213	0,231	0,223
HBB	0,051	0,139	0,212	0,180	0,216	0,184
SBB		0,163	0,171	0,257	0,169	**0,269**

Table 4.2: A - MRR using Full Image for g.

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,156	0,191	0,205	0,206	0,215
HBB	0,051	0,130	0,212	0,170	0,216	0,174
SBB		0,162	0,160	0,240	0,161	**0,258**

Table 4.3: A - MRR using Center Bias for g.

Diversity helps

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,150	0,240	0,274	0,249	**0,283**
HBB	0,051	0,173	0,200	0,136	0,206	0,147
SBB		0,178	0,168	0,242	0,174	0,257

Table 4.4: A - MRR using Saliency Maps for g.

Discussion & Conclusions

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I [5]	TVSS+I
FI		0,157	0,216	0,213	0,231	0,223
HBB	0,051	0,139	0,212	0,180	0,216	0,184
SBB		0,163	0,171	0,257	0,169	**0,269**

Table 4.2: A - MRR using Full Image for g.

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,156	0,191	0,205	0,206	0,215
HBB	0,051	0,130	0,212	0,170	0,216	0,174
SBB		0,162	0,160	0,240	0,161	**0,258**

Table 4.3: A - MRR using Center Bias for g.

Objects are not always in the center

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,150	0,240	0,274	0,249	**0,283**
HBB	0,051	0,173	0,200	0,136	0,206	0,147
SBB		0,178	0,168	0,242	0,174	0,257

Table 4.4: A - MRR using Saliency Maps for g.

Discussion & Conclusions

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I [5]	TVSS+I
FI		0,157	0,216	0,213	0,231	0,223
HBB	0,051	0,139	0,212	0,180	0,216	0,184
SBB		0,163	0,171	0,257	0,169	**0,269**

Table 4.2: A-MRR using Full Image for g.

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,156	0,191	0,205	0,206	0,215
HBB	0,051	0,130	0,212	0,170	0,216	0,174
SBB		0,162	0,160	0,240	0,161	**0,258**

Table 4.3: A-MRR using Center Bias for g.

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,150	0,240	0,274	0,249	**0,283**
HBB	0,051	0,173	0,200	0,136	0,206	0,147
SBB		0,178	0,168	0,242	0,174	0,257

Table 4.4: A-MRR using Saliency Maps for g.

Saliency Maps do not help here

But they do here

Discussion & Conclusions

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I [5]	TVSS+I
FI		0,157	0,216	0,213	0,231	0,223
HBB	0,051	0,139	0,212	0,180	0,216	0,184
SBB		0,163	0,171	0,257	0,169	**0,269**

Table 4.2: A - MRR using Full Image for g.

Parameters are **not independent**.

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,156	0,191	0,205	0,206	0,215
HBB	0,051	0,130	0,212	0,170	0,216	0,174
SBB		0,162	0,160	0,240	0,161	**0,258**

Table 4.3: A - MRR using Center Bias for g.

$f(Q)$	Time Sorting	Visual Ranking	NNDR	TVSS	NNDR+I	TVSS+I
FI		0,150	0,240	0,274	0,249	**0,283**
HBB	0,051	0,173	0,200	0,136	0,206	0,147
SBB		0,178	0,168	0,242	0,174	0,257

Table 4.4: A - MRR using Saliency Maps for g.

Where did I leave my phone ?

Cristian Reyes, Eva Mohedano, Kevin McGuinness and Noel E. O'Connor
Insight Centre for Data Analytics
Dublin, Ireland
cristian.reyes@estudiant.upc.edu
eva.mohedano@insight-centre.org

Xavier Giro-i-Nieto
Universitat Politecnica de Catalunya
Barcelona, Catalonia/Spain
xavier.giro@upc.edu

1. Introduction

The interest of users in having their lives digitally recorded has grown in the last years thanks to the advances on wearable sensors. Wearable cameras are one of the most informative ones, but they generate large amounts of images that require automatic analysis to build useful applications upon them. In this work we explore the potential of these devices to find the last appearance of personal objects among the more than 2,000 images that are generated everyday. This application could help into developing personal assistants capable of helping users when they do not remember where they left their personal objects. We adapt a previous work on instance search [3] to the specific domain of egocentric vision.

2. Methodology

Our goal is to rank the egocentric images captured during a day based on their likelihood to depict the location of a personal object. The whole pipeline is composed of the following stages: ranking by visual similarity, partition between candidate/non-candidate images and temporal-aware reranking within each class.

2.1. Ranking by Visual similarity

Given a certain set of query images Q depicting the object to be found, the algorithm starts by producing a ranking of the images of the day I ordered by their visual similarity score ν. This score is computed according to [3], which uses a bag of visual words model built with local features from a convolutional neural network (CNN).

A feature vector $q = f(Q)$ is generated from the set of images in Q that depict the object to locate. Three different approaches have been explored to define f:

a) No Mask: The q vector is built by averaging the visual words of all the local CNN features from the query images.

b) Mask: The q vector is built by averaging the visual words of the local CNN features that fall inside a query bounding box that surrounds the object. This allows to consider only the visual words that describe the object.

c) Weighted Mask: The q vector is built by averaging the visual words of the local CNN features of the whole image, but this time weighted depending on their distance to the bounding box. This allows to consider the context in addition to the object.

2.2. Detection of Candidate Moments

As a second step, a thresholding technique is applied to the ranking in order to partition the I set into two subsets named Candidates (C) and Discarded (D) moments.

Two different thresholding techniques were considered in order to create the C and $D = I \setminus C$ sets: TVSS (Threshold on Visual Similarity Scores) and NNDR (Nearest Neighbor Distance Ratio). The TVSS technique builds $C = \{i \in I : \nu_i > \nu_{th}\}$. The NNDR technique is based in the one described by Loewe [2]. Let ν_1 and ν_2 be the two best scores, then it builds $C = \left\{i \in I : \frac{\nu_i}{\nu_1} > p_{th}\frac{\nu_2}{\nu_1}\right\}$.

2.3. Temporal-aware reranking

The temporal-aware reranking step introduces the concept that the lost object is not in the location with the best visual match with the query, but in the last location where it was seen. Image sets R_C and R_D are built by reranking the elements in C and D, respectively, based on their time stamps. The final ranking R is built as the concatenation of $R = [R_C, R_D]$.

We considered two strategies for the temporal reranking: a straightforward sorting from the latest to the earliest timestamp, or a a more elaborate one that introduces diversity.

The diversity-aware configuration avoids presenting consecutive images of the same *moment* in the final ranked list. This is especially important in egocentric vision, where sequential images in time often present a high redundancy. Our diversity-based technique is based in the interleaving of samples, which is frequently used in digital communication. It consists in ordering temporally the images in I but knowing for each image if it belongs to C or D. So we might have something similar to $O = \{i_1^D, ..., i_{k-1}^D, i_k^C, ..., i_{l-1}^C, i_l^D, ..., i_{m-1}^D, i_m^C, ..., i_{n-1}^C\}$. Then $R_C = \{i_k^C, i_l^C, i_m^C, i_{k+1}^C, i_{l+1}^C, i_{m+1}^C, i_{k+2}^C, ...\}$ and R_D is built analogously.

3. Experiments

3.1. Dataset annotation

Our work has been developed over the NTCIR Lifelogging Dataset [1] which consists of anonymised images taken every 30 seconds over a period of 30 days. Each day contains around 1,500 images.

This dataset was annotated for this work with five personal objects which could be lost: a phone, headphones, a watch and a laptop. In particular, they were tagged as *relevant* the last appearance of the object within each day.

Queries were defined by considering that the user had a collection of images of the object, not only one. The Q set contained from 3 to 5 images per category. These images showed the objects clearly and were used to build the q vector. This assumption is realistic as the object to be found could be defined from past appearances from the same dataset.

3.2. Training

The proposed system presents some parameters that were learned with the training part of the dataset.

A visual vocabulary for Bag of Words was learned from around 14,000 images of 9 days, generating a total of 25,000 centroids. The thresholds ν_{th} and p_{th} respectively were also learned on the same 9 days used for training. The optimal values found are detailed in Table 1.

	No Mask	Mask	Weighted Mask
ν_{th}	0.04	0.01	**0.04**
p_{th}	0.17	**0.11**	0.14

Table 1. Optimal thresholds. In bold those that gave highest mAP

3.3. Test

For evaluating the performance, Mean Average Precision (mAP) was computed for each day, taking into account all the categories. Then these values have been averaged over 15 test days and presented in Table 2.

Applying a thresholding technique has demonstrated to be helpful, as the combination of the object masking and the NNDR thresholding technique has shown the best results.

It must be noticed that mAP is not the best measure in diversity terms, so despite the fact that mAP decreases, the images that form the top of the ranking have shown to be from more diverse scenes as it is shown in Figure 1.

Figure 1. Results obtained for a search in category phone for a certain day. First row are the images that form Q with mask, second row results using NNDR and third results using NNDR + Div.

	No Mask	Mask	Weighted Mask
Temporal Ordering	0.051	0.051	0.051
Visual Similarity	0.102	0.082	0.111
TVSS	0.113	0.111	0.139
NNDR	0.086	**0.176**	0.093
TVSS + Div	0.096	0.082	0.118
NNDR + Div	0.066	**0.166**	0.049

Table 2. mAP results obtained when testing over 15 days.

4. Conclusions

This work has presented a good baseline for further research on the problem of finding the last appearance of an object in egocentric images.

Instance search based on bags of convolutional local features has shown promising results on egocentric images. Thresholding and temporal diversity techniques have improved the performance of visual only cues.

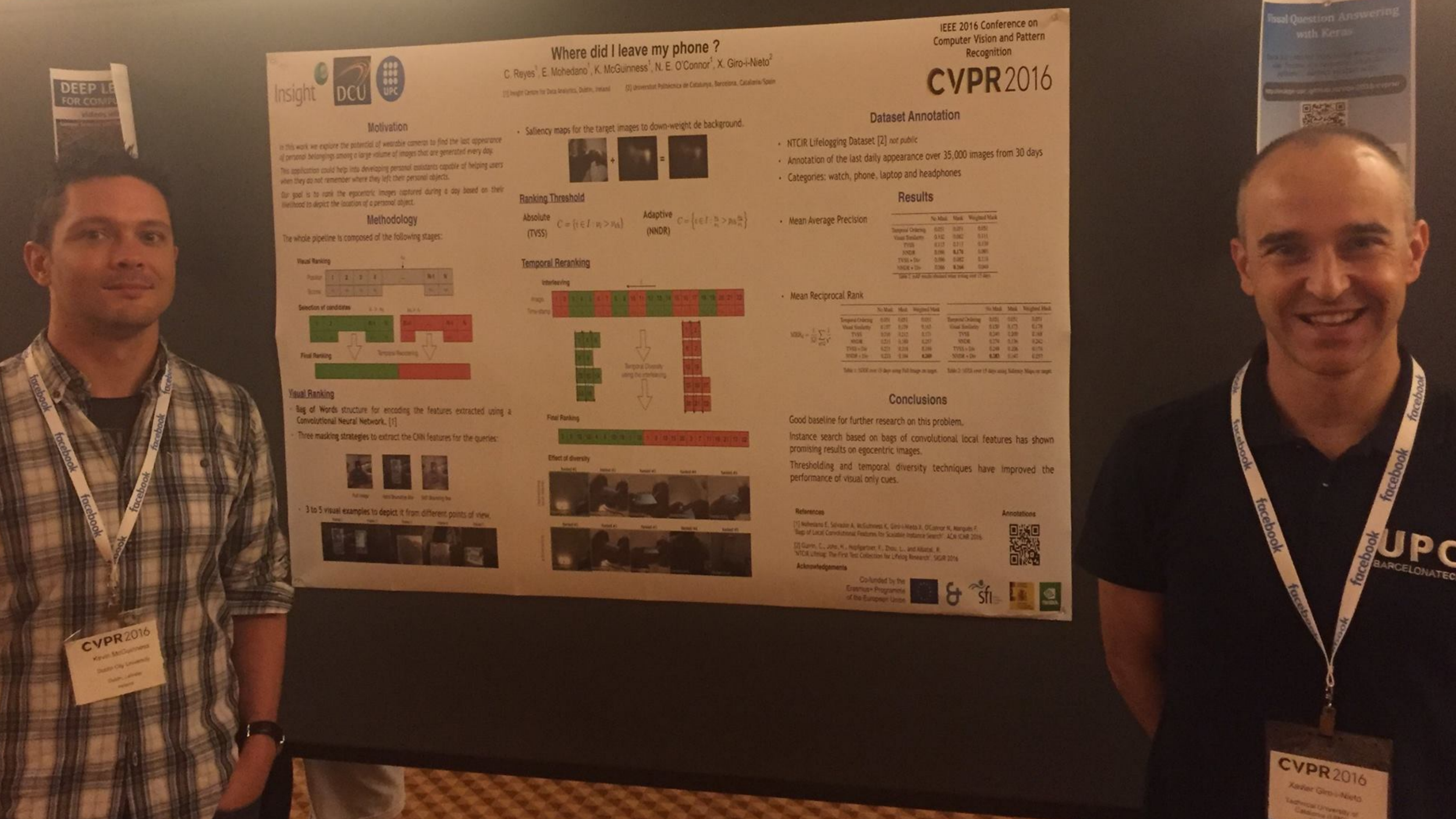
We plan to extend the annotations to neighbor images that may also depict relevant information to locate the location where the object was found. This way, not only one image would be considered as relevant, as assumed in the presented experiments.

References

[1] C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, and R. Albatal. Ntcir lifelog: The first test collection for lifelog research. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Pisa, Italy, July 2016. ACM.

[2] D. G. Loewe. Distinctive image features from scale-invariant keypoints. page 20, 2004.

[3] E. Mohedano, A. Salvador, K. McGuinness, F. Marques, N. E. O'Connor, and X. Giro-i Nieto. Bags of local convolutional features for scalable instance search. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 2016.



Where did I leave my phone ?
C. Reyes, E. Mohedano, K. McGuinness, N. E. O'Connor, X. Giro-i-Nieto
IEEE 2016 Conference on Computer Vision and Pattern Recognition
CVPR2016
Insight
DCU
UPC
Motivation
Methodology
The whole pipeline is composed of the following stages:
Visual Ranking
Selection of candidates
Final Ranking
Visual Ranking
Bag of Words structure for encoding the features extracted using a Convolutional Neural Network. [1]
Three masking strategies to extract the CNN features for the queries:
3 to 5 visual examples to depict it from different points of view.
Saliency maps for the target images to down-weight de background.
Ranking Threshold
Absolute (TVSS)
Adaptive (NNDR)
Temporal Reranking
Interleaving
Final Ranking
Effect of diversity
Dataset Annotation
NTCIR Lifelogging Dataset [2] not public
Annotation of the last daily appearance over 35,000 images from 30 days
Categories: watch, phone, laptop and headphones
Results
Mean Average Precision
Mean Reciprocal Rank
Conclusions
Good baseline for further research on this problem.
Instance search based on bags of convolutional local features has shown promising results on egocentric images.
Thresholding and temporal diversity techniques have improved the performance of visual only cues.
References
Annotations
Acknowledgements
CVPR2016
CVPR2016
DEEP LE
FOR COMP
facebook
UPC
BARCELONATEC

Lifelogging Tools and Applications

ACM multimedia

Acknowledgements

Noel E. O'Connor

Cathal Gurrin

Albert Gil

Financial Support

sfi Science Foundation Ireland

Generalitat de Catalunya
Departament d'Economia
i Coneixement

Erasmus+


THANK YOU

f(Q)	NNDR	TVSS	NNDR + I	TVSS + I
Saliency Mask	0,176	0,271	0,190	**0,279**

Using Full Image for target

f(Q)	NNDR	TVSS	NNDR + I	TVSS + I
Saliency Mask	0,162	0,198	0,173	**0,213**

Using Center Bias for target

f(Q)	NNDR	TVSS	NNDR + I	TVSS + I
Saliency Mask	0,201	0,217	0,210	**0,226**

Using Saliency Mask for target

Visual Ranking - Descriptors

Bag of Words

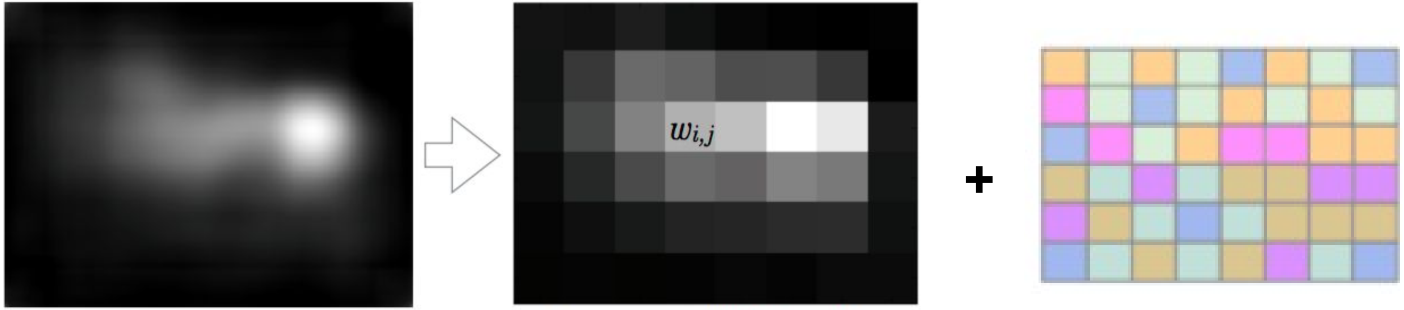
An example using text

A) There is a red ball in a white box.

B) The red box contains a white ball.

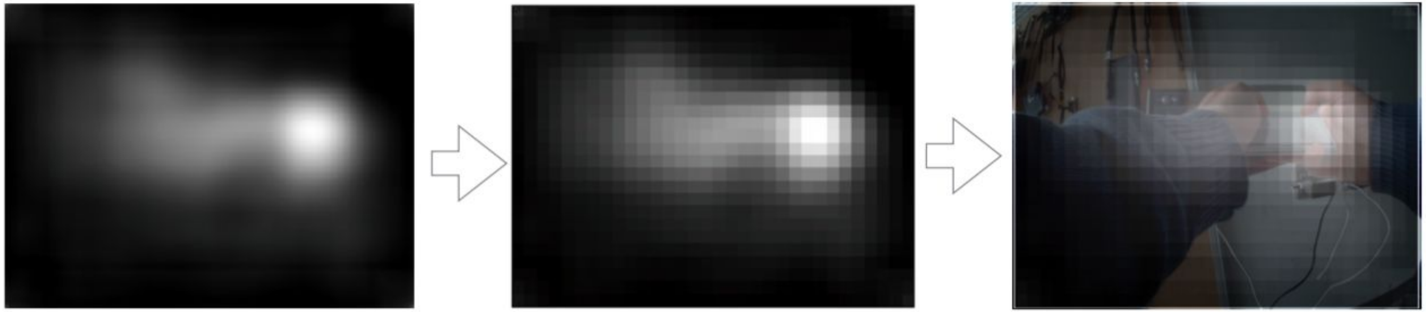
Cosine Similarity Score = **0.68**

	Sentence A	Sentence B
there	1	0
is	1	0
a	2	1
red	1	1
ball	1	1
in	1	0
white	1	1
box	1	1
the	0	1
contains	0	1



Original Saliency Map

6x8 down-sampled Saliency Map



Original Saliency Map

32x42 down-sampled Saliency Map

Effect of weighting with the saliency map

Outline

1. Motivation
2. Methodology
3. Experiments
4. Results
5. **Conclusions**

Conclusions

- The **system** accomplishes its task.
- **Thresholding** and **temporal reranking** have improved performance
- **Center Bias** does not necessary improve performance.
- **Saliency Maps** have improved performance.
- **Parameters** are **not independent** when measuring with A-MRR.
- Good **baseline** for further research.