LIFELOGGING TOOLS & APPLICATIONS WORKSHOP

ACM MULTIMEDIA

Mountain View, CA, USA
23 October 2017

[Slides on Slideshare] @DocXavi

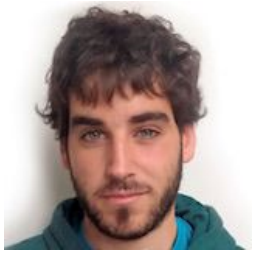
#ACMMM
#lifelogging

Semantic Summarization of Egocentric Photo Stream Events

Xavier Giro-i-Nieto
xavier.giro@upc.edu

Associate Professor
Universitat Politecnica de Catalunya
Technical University of Catalonia

Joint Work



Aniol Lidon

Marc Bolaños

Mariella Dimiccoli

Petia Radeva

Maite Garolera

Xavi Giró

UPC

Image Processing Group

Universitat de Barcelona

Barcelona Perceptual Computing Laboratory

CST

Consorci Sanitari de Terrassa

Brain, Cognition and Behaviour Group

Motivation

- In 2013, 44.4 million people with dementia worldwide.
- “Cognitive Stimulation Therapy”

Motivation

Use episodic images to develop cognitive exercises and tools for memory reinforcing of MCI and Alzheimer patients.

Motivation

Egocentric photo streams captured from a wearable camera.

P. Piasek, K. Irving, and A.F. Smeaton, “SenseCam intervention based on Cognitive Stimulation Therapy framework for early-stage dementia”. In Pervasive Computing Technologies for Healthcare 2011.

Motivation

- Low temporal resolution (2 frames per minute on average).
- Up to 2000~3000 images at day!
- **Summarization** is needed.

Overview

Automatically summarize already detected events.

- Ranking by semantic relevance.

RELEVANCE

Overview

Automatically summarize already detected events.

- Ranking by semantic relevance.
- Reranking based on novelty (visual diversity).

Overview

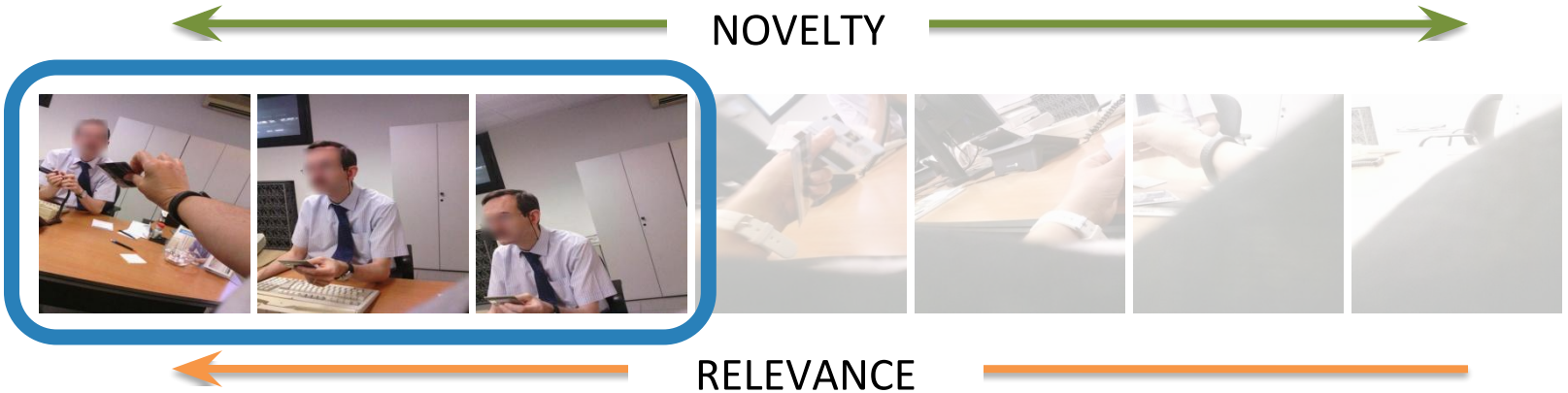
Automatically summarize already detected events.

- Ranking by semantic relevance.
- Reranking based on novelty (visual diversity).
- Image sets built from the top-T ranked images.

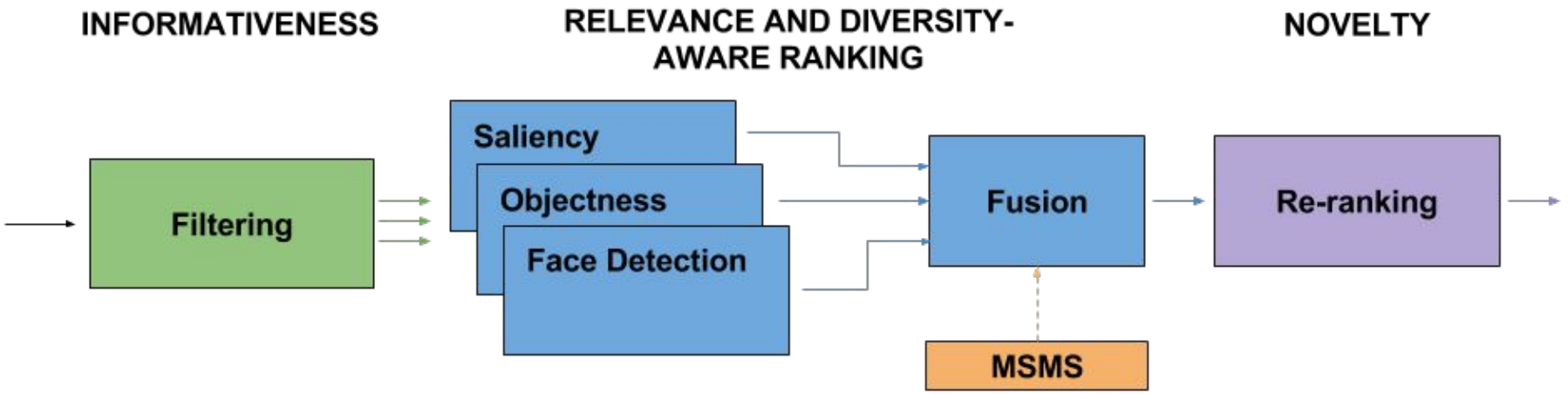
Overview

Automatically summarize already detected events.

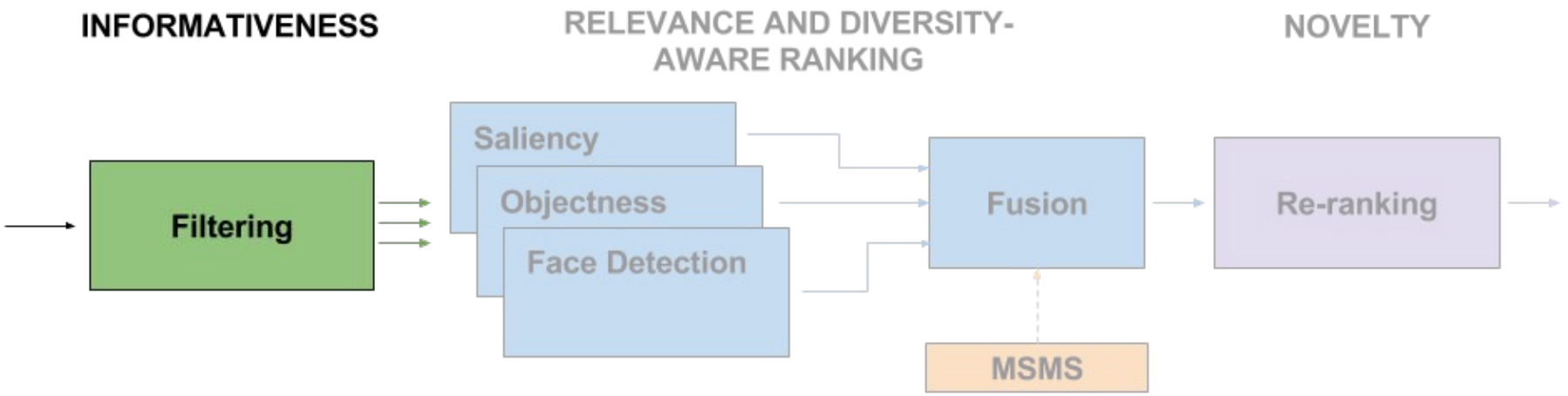
- Ranking by semantic relevance.
- Reranking based on novelty (visual diversity).
- Image sets built from the top-T ranked images.
- Final summary preserves the original temporal ordering.



System Architecture



System Architecture

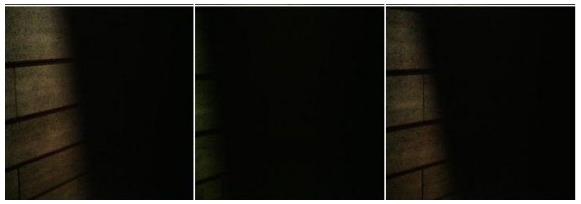


Informativeness Filtering

Aim: Removal of uninformative images, such as.

Blur

Black



Burned

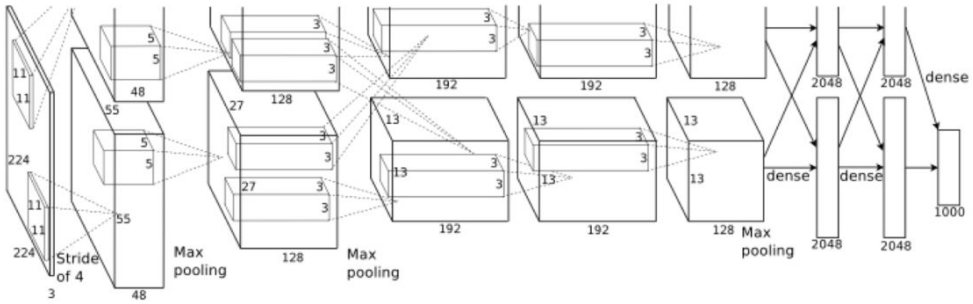
..and

- sky
- ceilings
- walls
- large occlusions

...

Informativeness Filtering

Fine-tunning an AlexNet-like CNN (CaffeNet) for binary classification:



IMAGENET

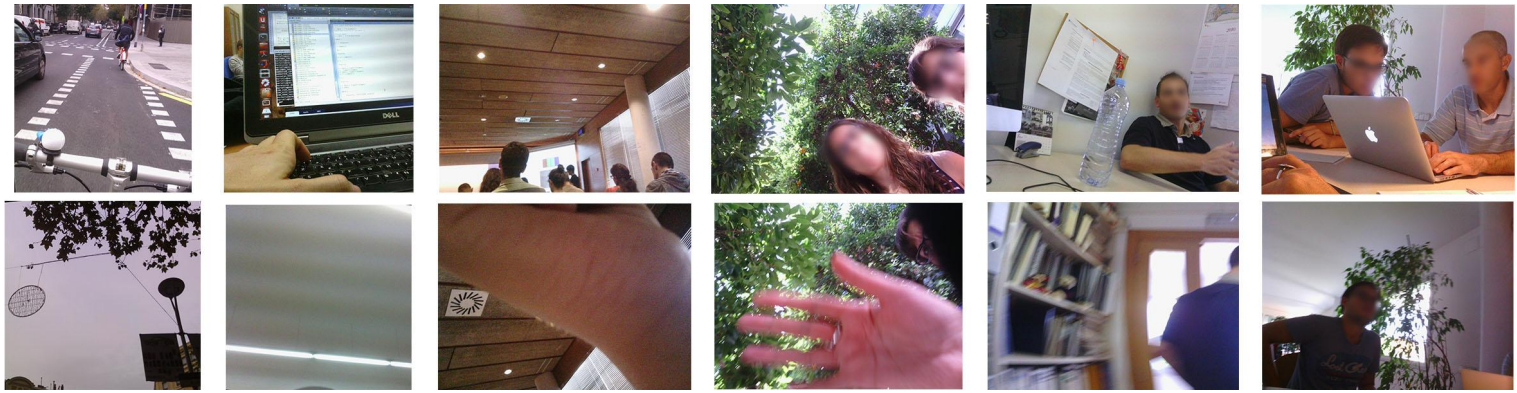
Informative

Non-Informative

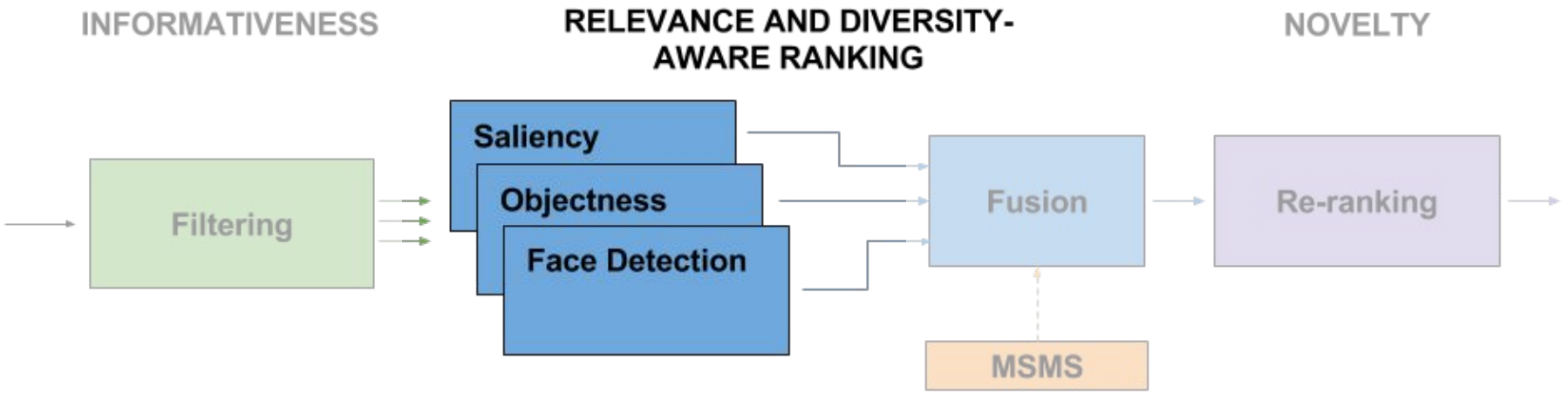
Informativeness Filtering

Informative

Non-Informative



System Architecture



Relevance

What is relevant ?

Frame-level:

- **Repeated**.
- **Unusual**.
- WHAT? **Representative** of an activity.
- WHO? **Social interactions**.
- WHERE? **Environment**.
- **WHEN** an event has occurred.
- **HOW** activity occurred.

Relevance

What is relevant ?

Frame-level:

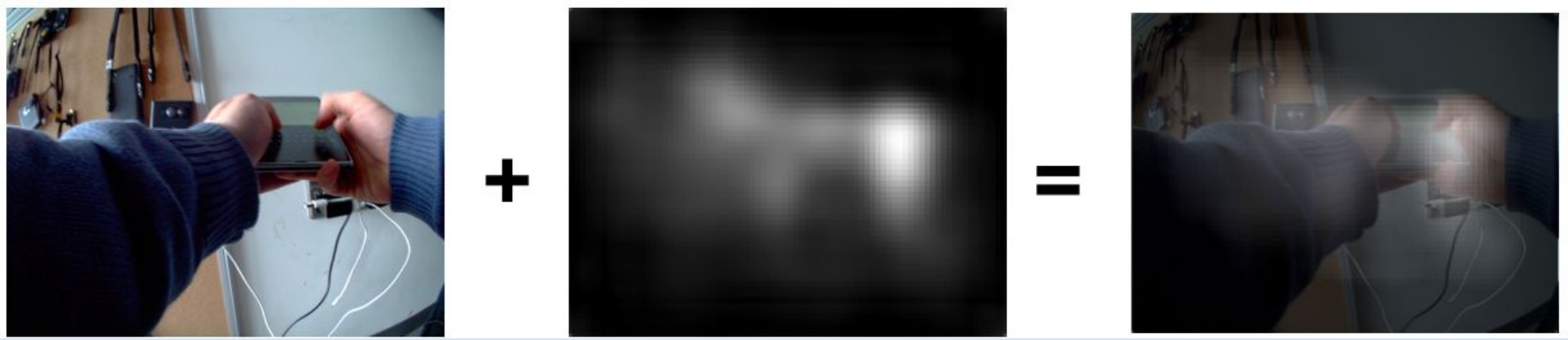
- WHAT? **Representative** of an activity.
  - Saliency Maps
  - Object detection

- WHO? **Social interactions**.
  - Face detection

Relevance ranking

Saliency maps

Aim: Determining interesting zones.



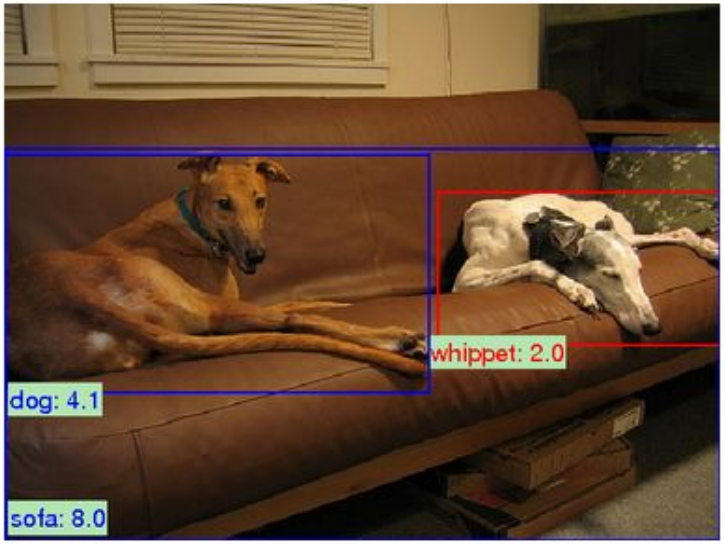
Relevance score: Average saliency per image.

Junting Pan, Kevin McGuinness, Elisa Sayorl, Noel E. O'Connor, Xavier Giró-i-Nieto, "Shallow and Deep Convolutional Networks for Saliency Prediction". CVPR 2016.

Relevance ranking

Objects

Aim: WHAT? **Representative** of an activity.



Relevance score: Sum of all object detection scores.

Judy Hoffman,Sergio Guadarrama, EricTzeng,Je Donahue,RossB.Girshick, Trevor Darrell, and Kate Saenko. 2014. LSDA: Large Scale Detection Through Adaptation. CoRR abs/1407.5035 (2014).

Relevance ranking

Faces

Aim: WHO? **Social interactions.**

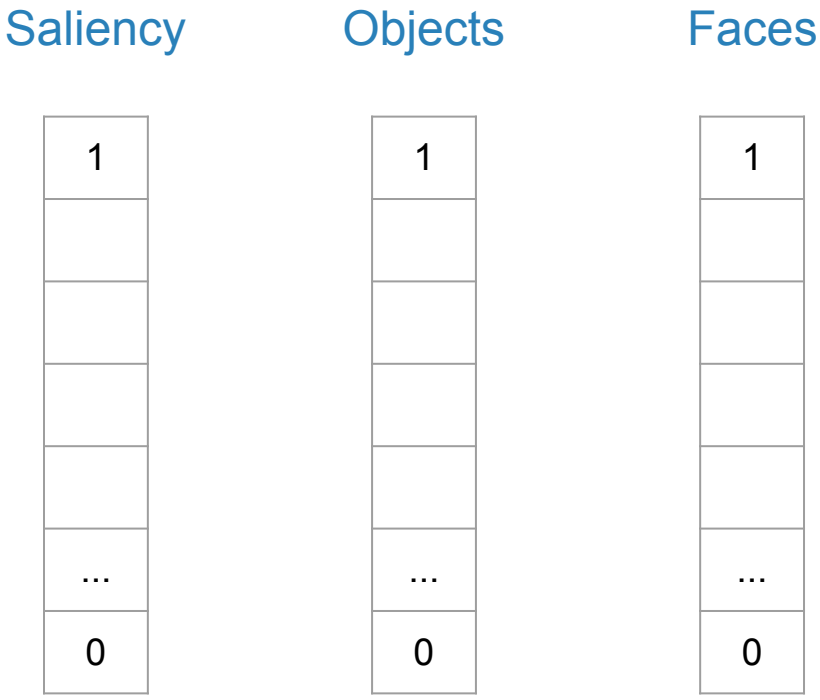
Relevance score: Exponential sum of face detection scores.

Xiangxin Zhu and Deva Ramanan. 2012. Face detection, pose estimation, and landmark localization in the wild. CVPR 2012.

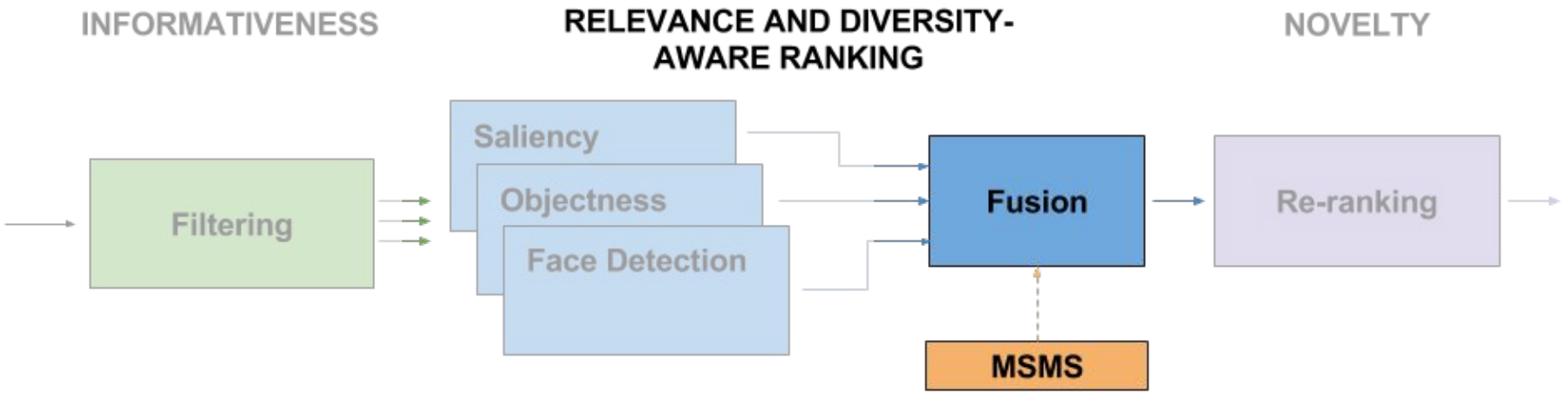
Relevance ranking

A ranked list $r_k(x)$ is built for each of the k semantic-aware rankings of M elements, with scores normalized by position.

$$r_k(x) = \frac{M - R_k(x)}{M - 1}$$

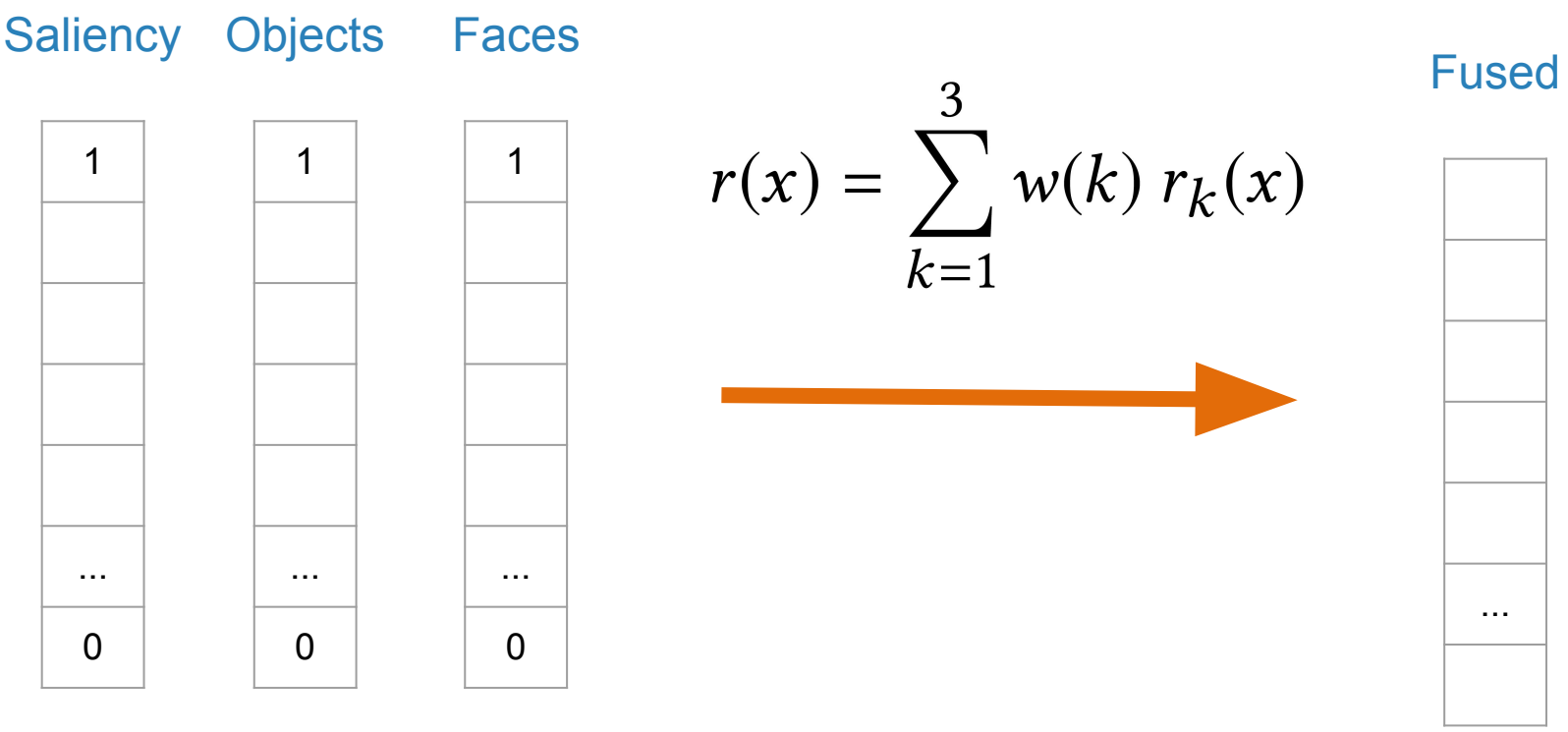


System Architecture



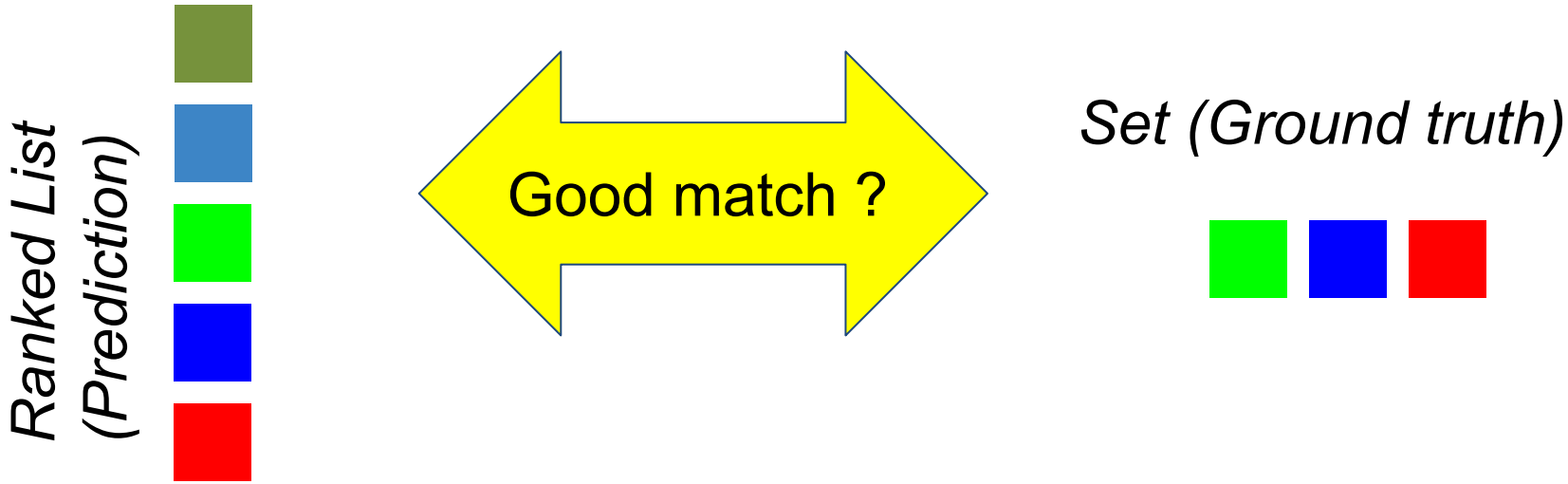
Weighted Fusion of Ranks

A weighted linear combination of scores to build r(x).



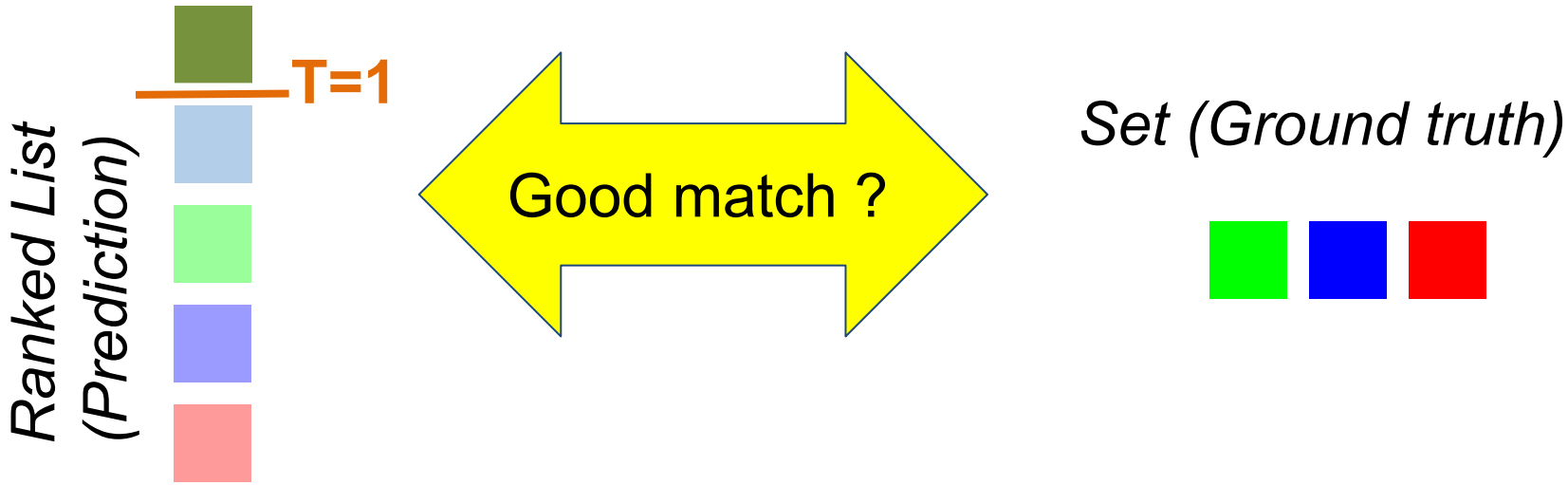
Weighted Fusion of Ranks

Weights can be estimated based on the performance of each relevance rank separately in solving the summarization task.



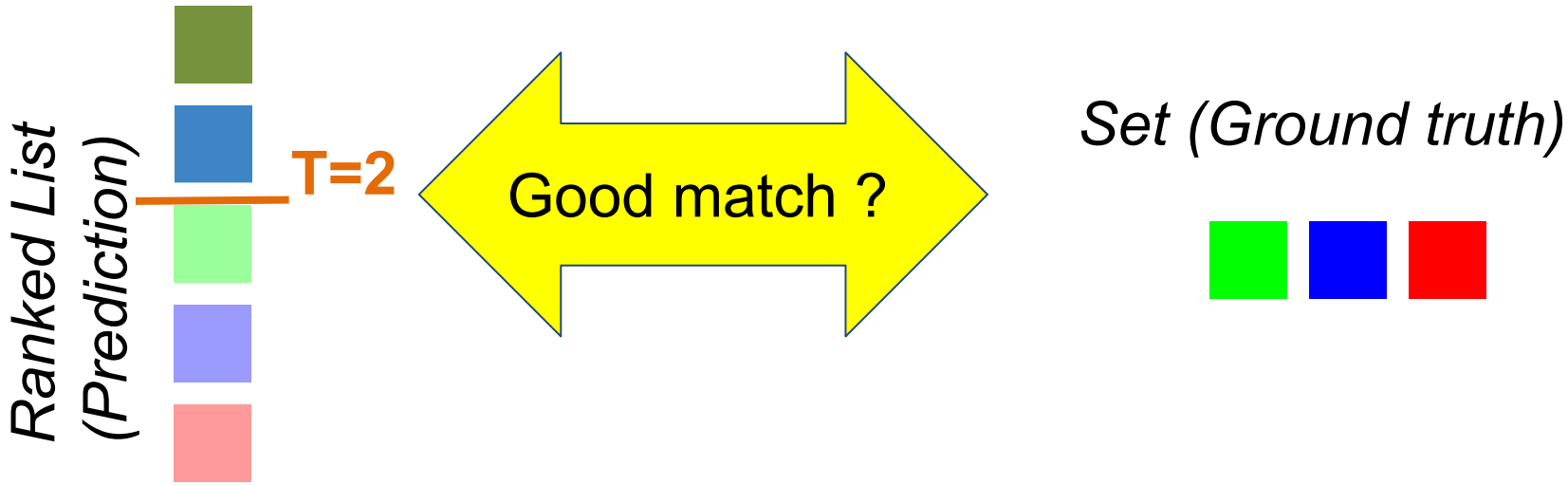
Weighted Fusion of Ranks

Summaries of any size (T) can be built from the ranked list by **thresholding**.



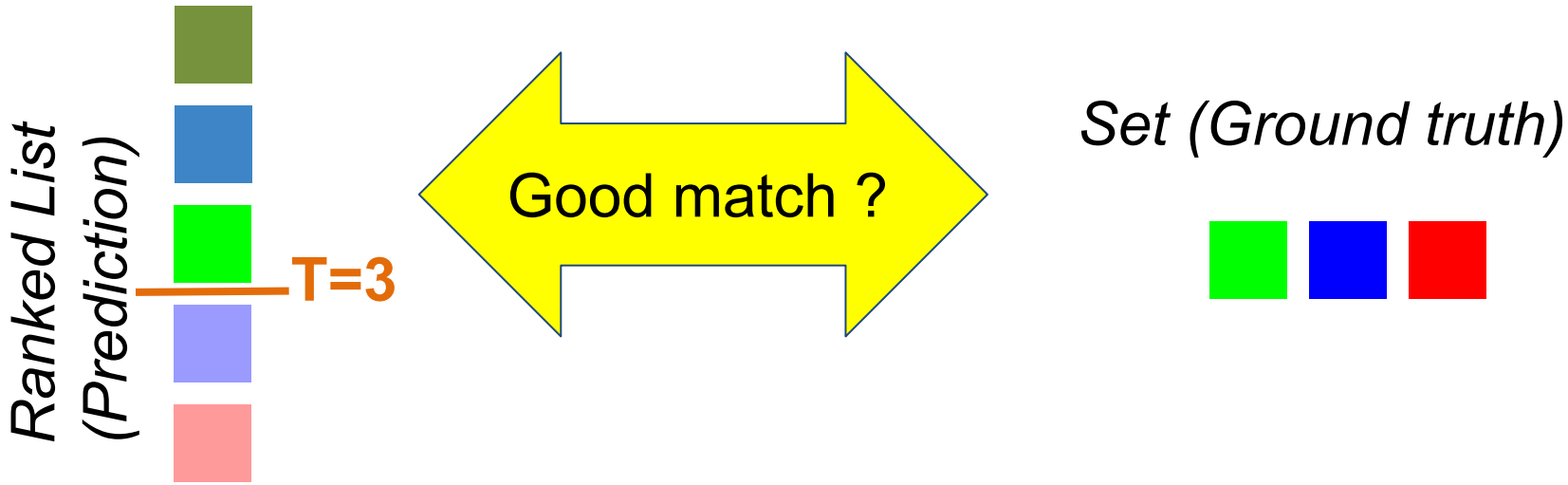
Weighted Fusion of Ranks

Summaries of any size (T) can be built from the ranked list by **thresholding**.



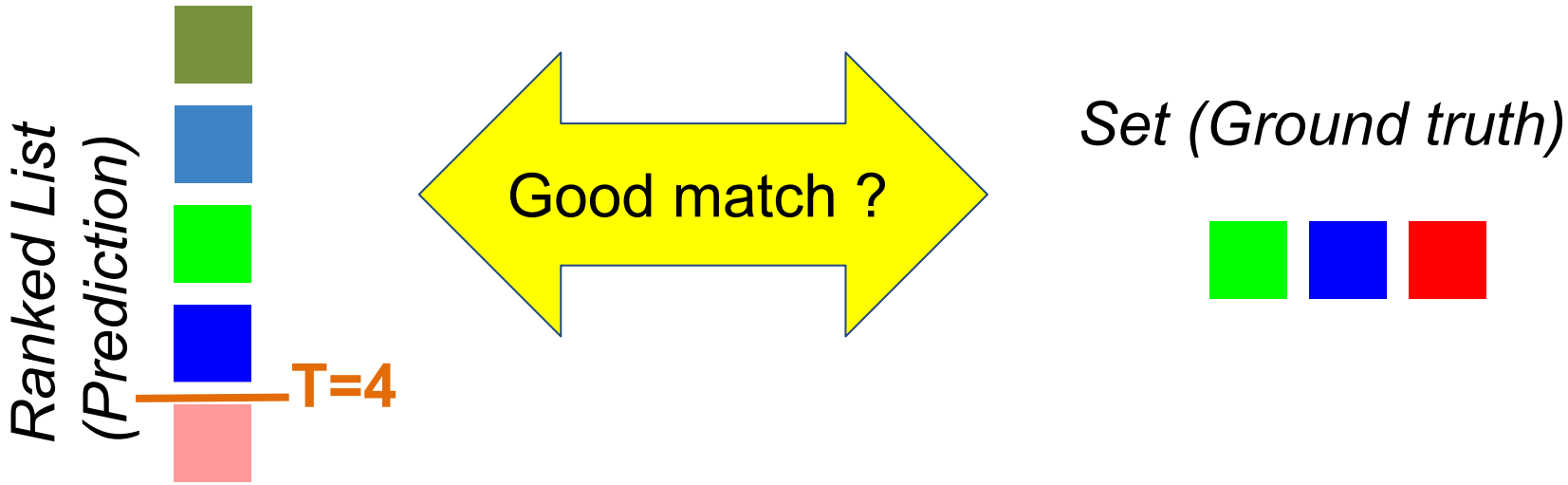
Weighted Fusion of Ranks

Summaries of any size (T) can be built from the ranked list by **thresholding**.



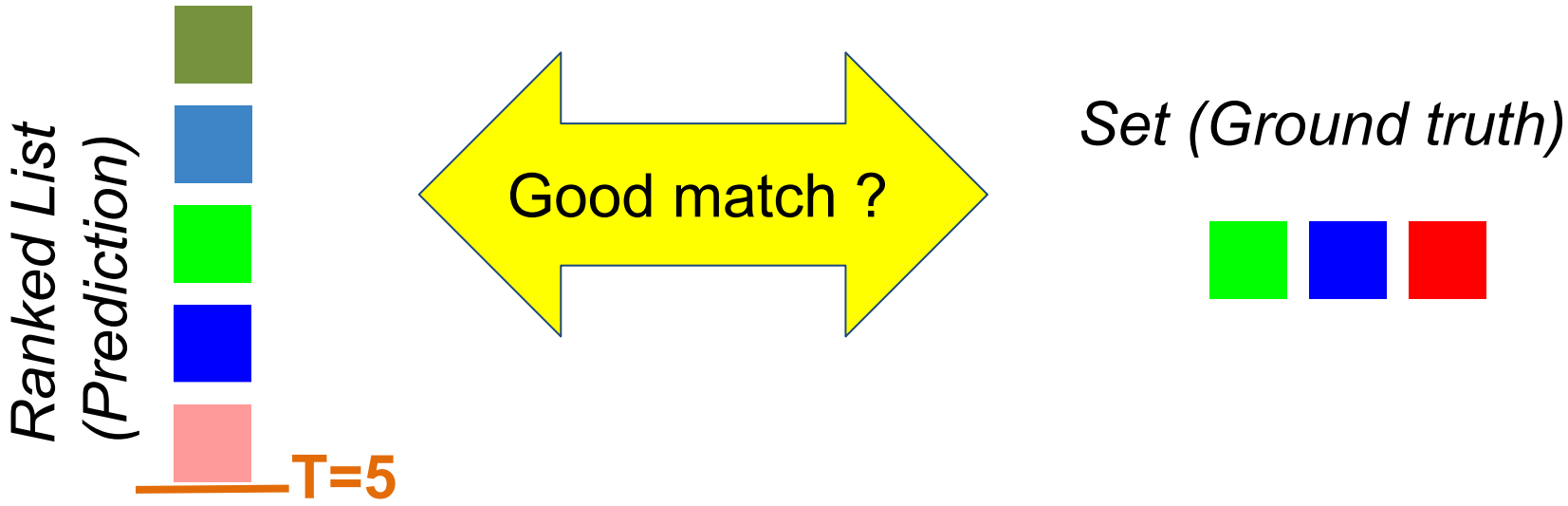
Weighted Fusion of Ranks

Summaries of any size (T) can be built from the ranked list by **thresholding**.



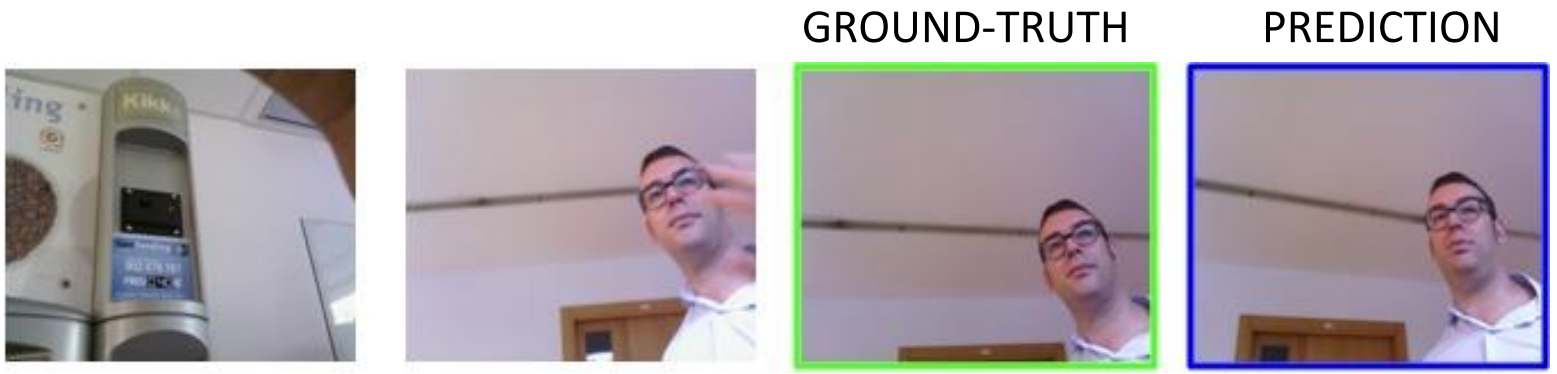
Weighted Fusion of Ranks

Summaries of any size (T) can be built from the ranked list by **thresholding**.

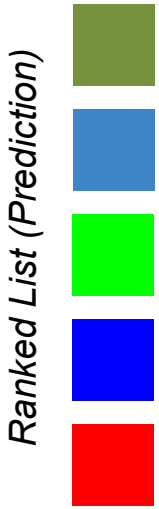


Weighted Fusion of Ranks

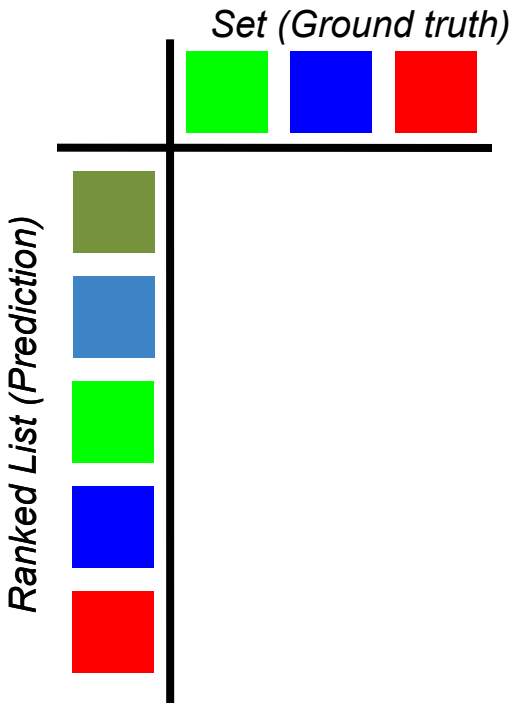
Forcing an exact matching between predictions and ground truth is a too hard approach.



Weighted Fusion of Ranks

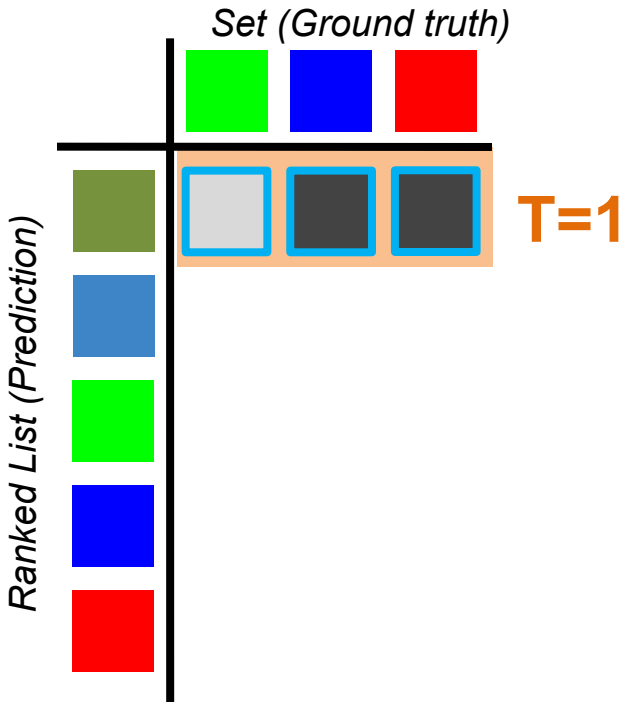


Weighted Fusion of Ranks



Weighted Fusion of Ranks

[Averaged] Sum of Max Similarities (SMS):



$$SMS(\mathcal{V}, \mathcal{Y}^T) = \frac{1}{P} \sum_{i=1}^{P} s(x_{v_i}, \mathcal{Y}^T)$$

Similarity Sum=  + +

P: Size of ground truth summary

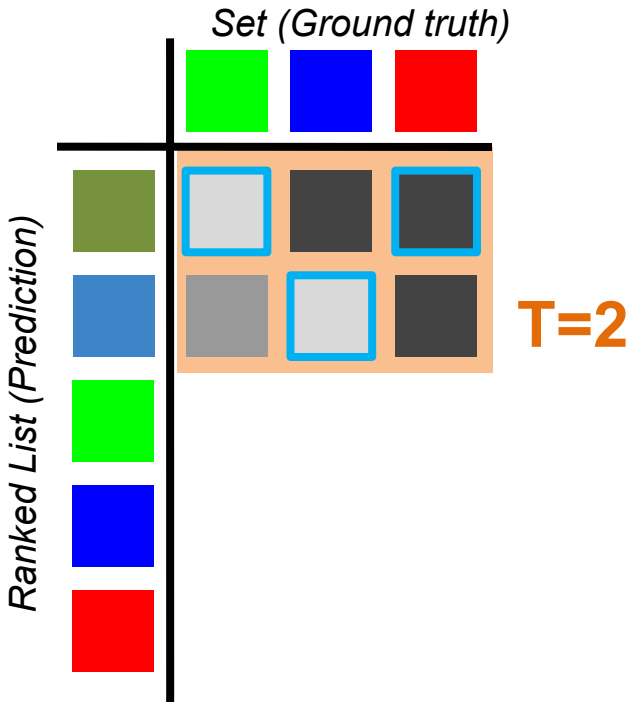
x_{vi}: Images in the ground truth set

Y^T: Predicted summary of size T

$s(x,Y^T)$: Maximum similarity of x with respect to all elements in set Y

Weighted Fusion of Ranks

[Averaged] Sum of Max Similarities (SMS):



$$SMS(\mathcal{V}, \mathcal{Y}^T) = \frac{1}{P} \sum_{i=1}^{P} s(x_{v_i}, \mathcal{Y}^T)$$

P: Size of ground truth summary

x_{vi}: Images in the ground truth set

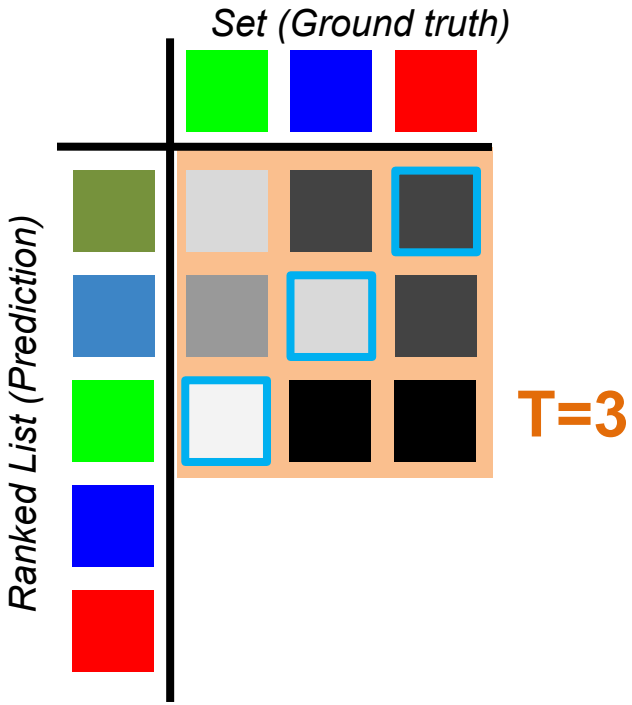
Y^T: Predicted summary of size T

$s(x,Y^T)$: Maximum similarity of x with respect to all elements in set Y

Similarity Sum= + +

Weighted Fusion of Ranks

[Averaged] Sum of Max Similarities (SMS):



$$SMS(\mathcal{V}, \mathcal{Y}^T) = \frac{1}{P} \sum_{i=1}^{P} s(x_{v_i}, \mathcal{Y}^T)$$

P: Size of ground truth summary

x_{vi}: Images in the ground truth set

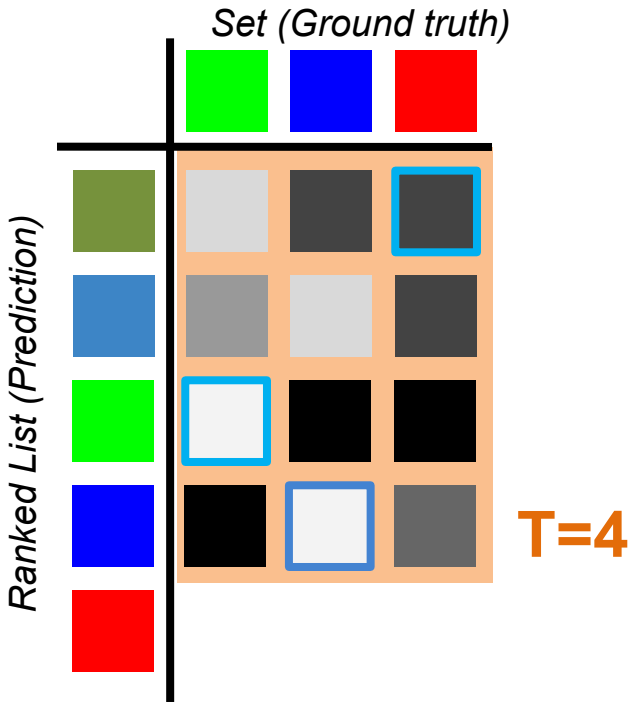
Y^T: Predicted summary of size T

$s(x,Y^T)$: Maximum similarity of x with respect to all elements in set Y

Similarity Sum= + +

Weighted Fusion of Ranks

[Averaged] Sum of Max Similarities (SMS):



$$SMS(\mathcal{V}, \mathcal{Y}^T) = \frac{1}{P} \sum_{i=1}^{P} s(x_{v_i}, \mathcal{Y}^T)$$

P: Size of ground truth summary

x_{vi}: Images in the ground truth set

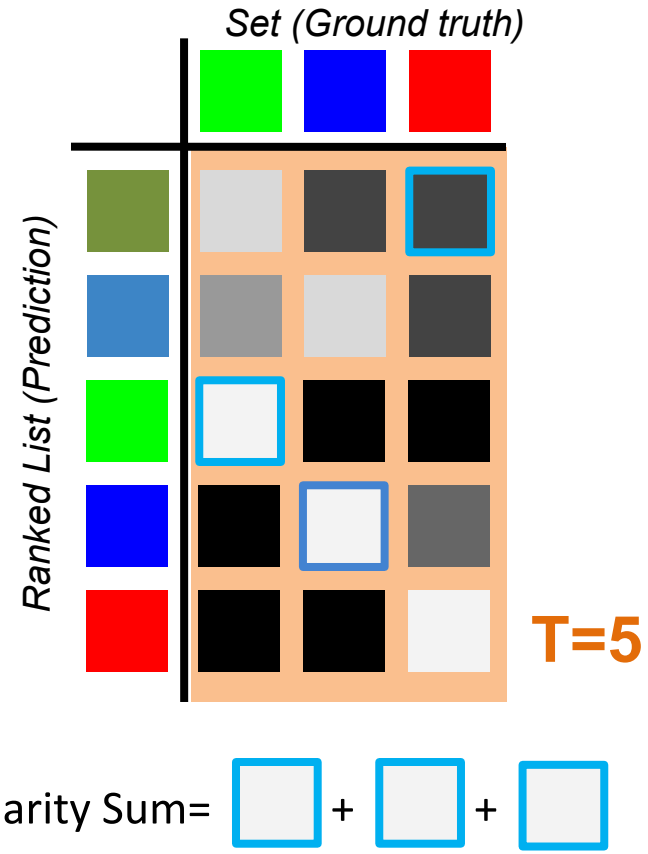
Y^T: Predicted summary of size T

$s(x,Y^T)$: Maximum similarity of x with respect to all elements in set Y

Similarity Sum= + +

Weighted Fusion of Ranks

[Averaged] Sum of Max Similarities (SMS):



$$SMS(\mathcal{V}, \mathcal{Y}^T) = \frac{1}{P} \sum_{i=1}^{P} s(x_{v_i}, \mathcal{Y}^T)$$

P: Size of ground truth summary

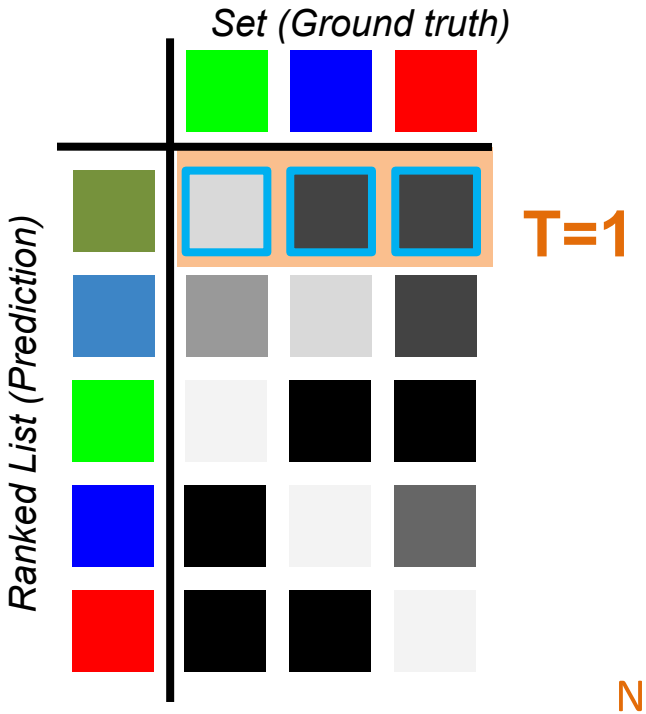
x_{vi}: Images in the ground truth set

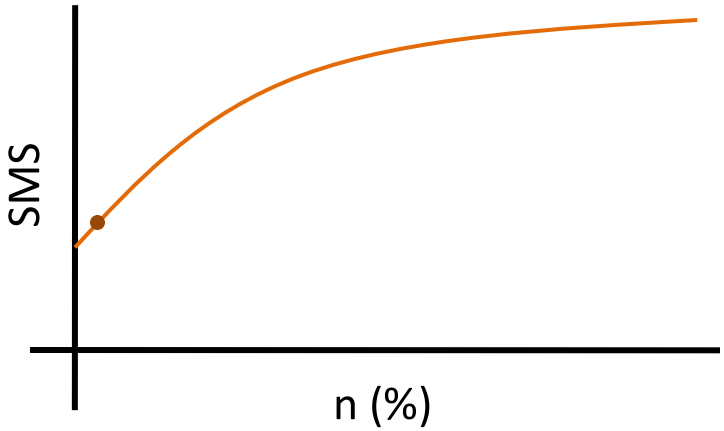
Y^T: Predicted summary of size T

$s(x,Y^T)$: Maximum similarity of x with respect to all elements in set Y

Weighted Fusion of Ranks

[Averaged] Sum of Max Similarities (SMS):





Similarity Sum= + +

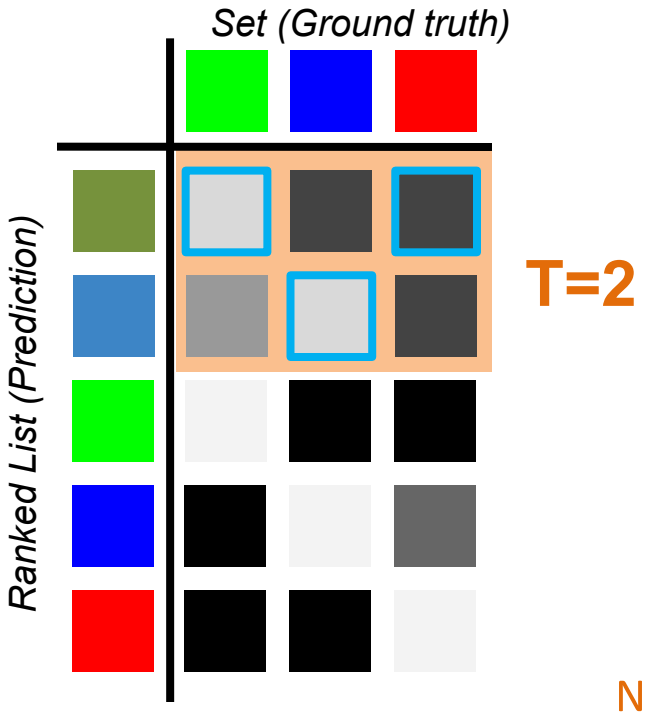
Normalization in both axes

Y: Divide by GT samples

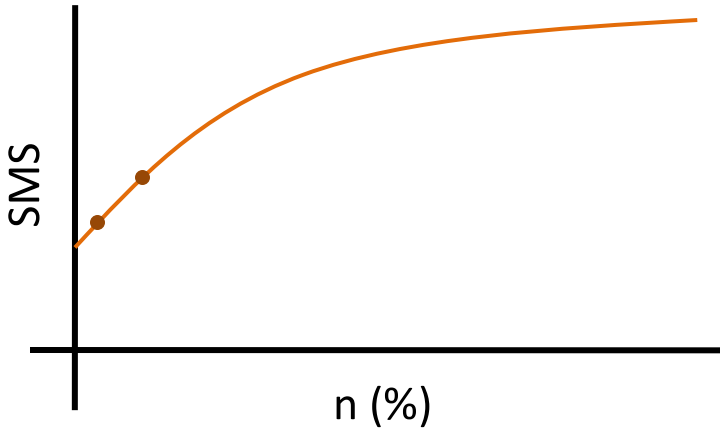
X: Reshape samples to N bins

Weighted Fusion of Ranks

[Averaged] Sum of Max Similarities (SMS):



T=2



Similarity Sum= + +

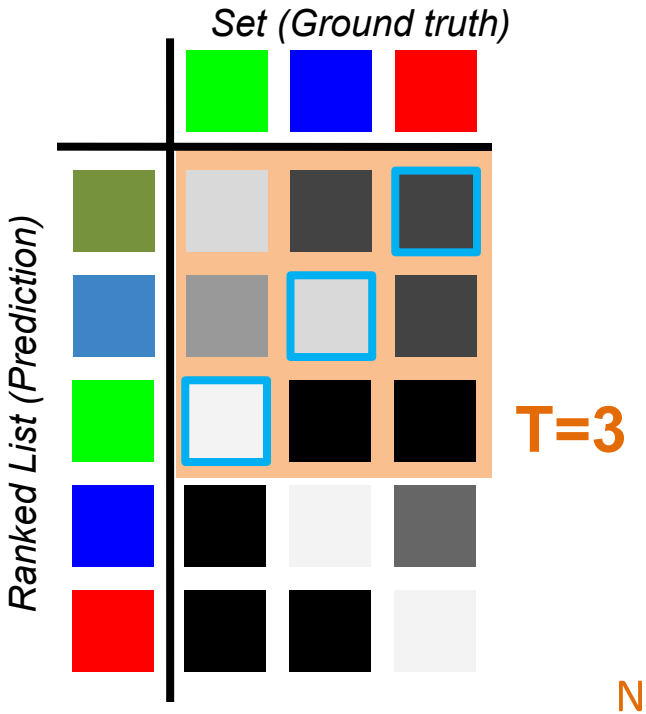
Normalization in both axes

Y: Divide by GT samples

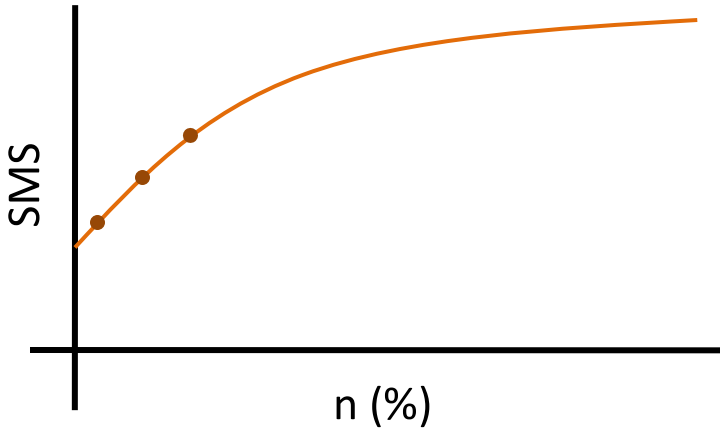
X: Reshape samples to N bins

Weighted Fusion of Ranks

[Averaged] Sum of Max Similarities (SMS):



T=3



Similarity Sum= + +

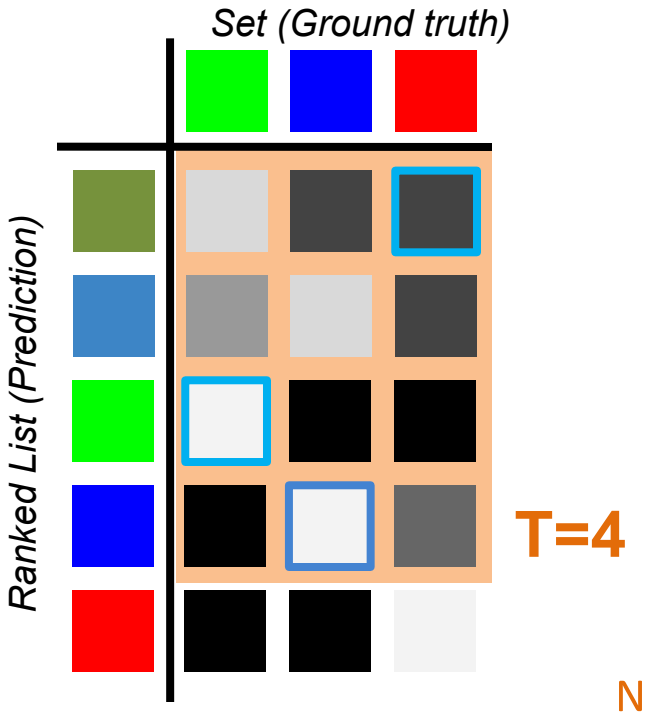
Normalization in both axes

Y: Divide by GT samples

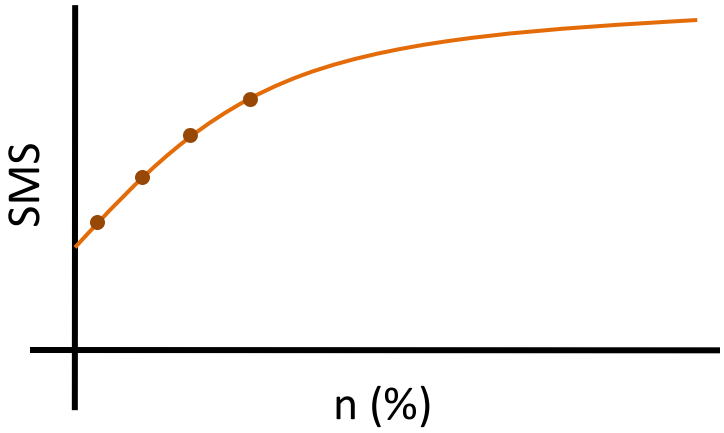
X: Reshape samples to N bins

Weighted Fusion of Ranks

[Averaged] Sum of Max Similarities (SMS):



T=4



Similarity Sum= + +

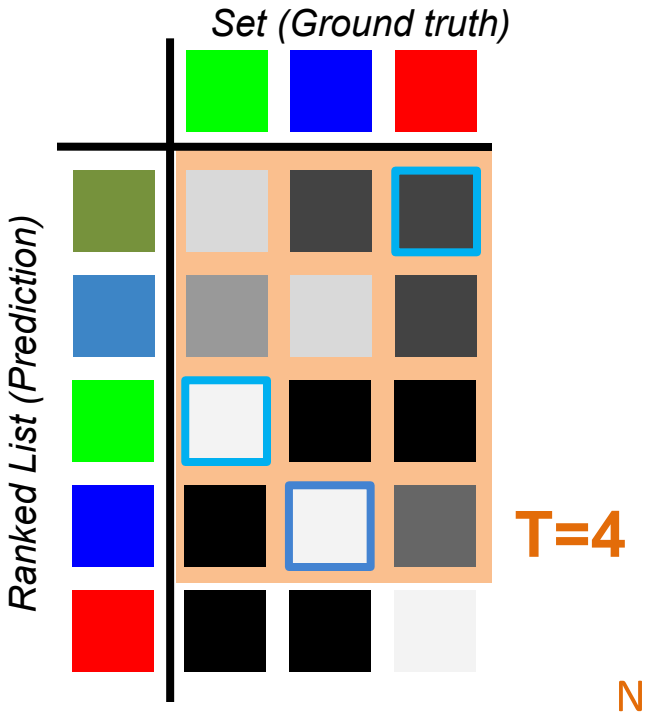
Normalization in both axes

Y: Divide by GT samples

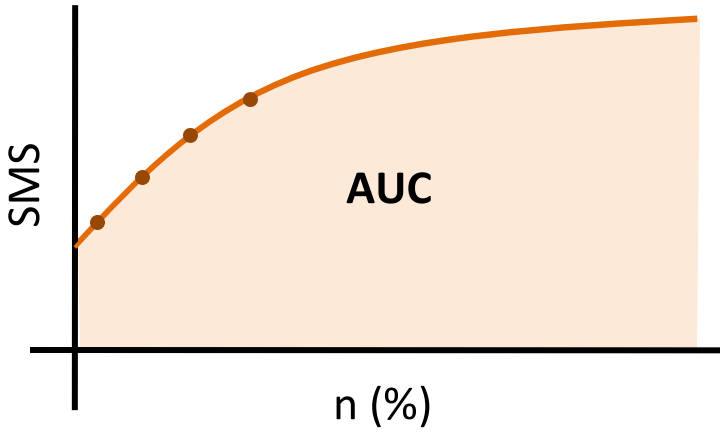
X: Reshape samples to N bins

Weighted Fusion of Ranks

[Averaged] Sum of Max Similarities (SMS):



T=4



Similarity Sum= + +

Normalization in both axes

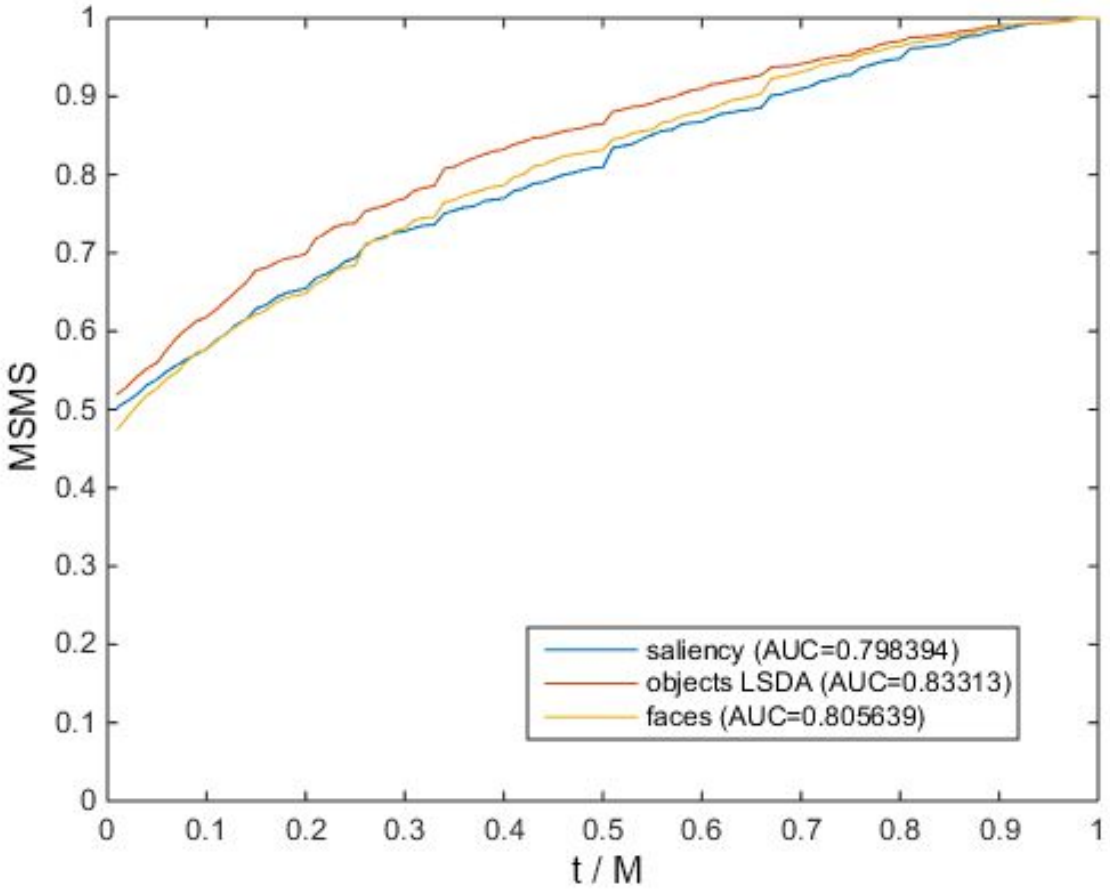
Y: Divide by GT samples

X: Reshape samples to N bins

Weighted Fusion of Ranks

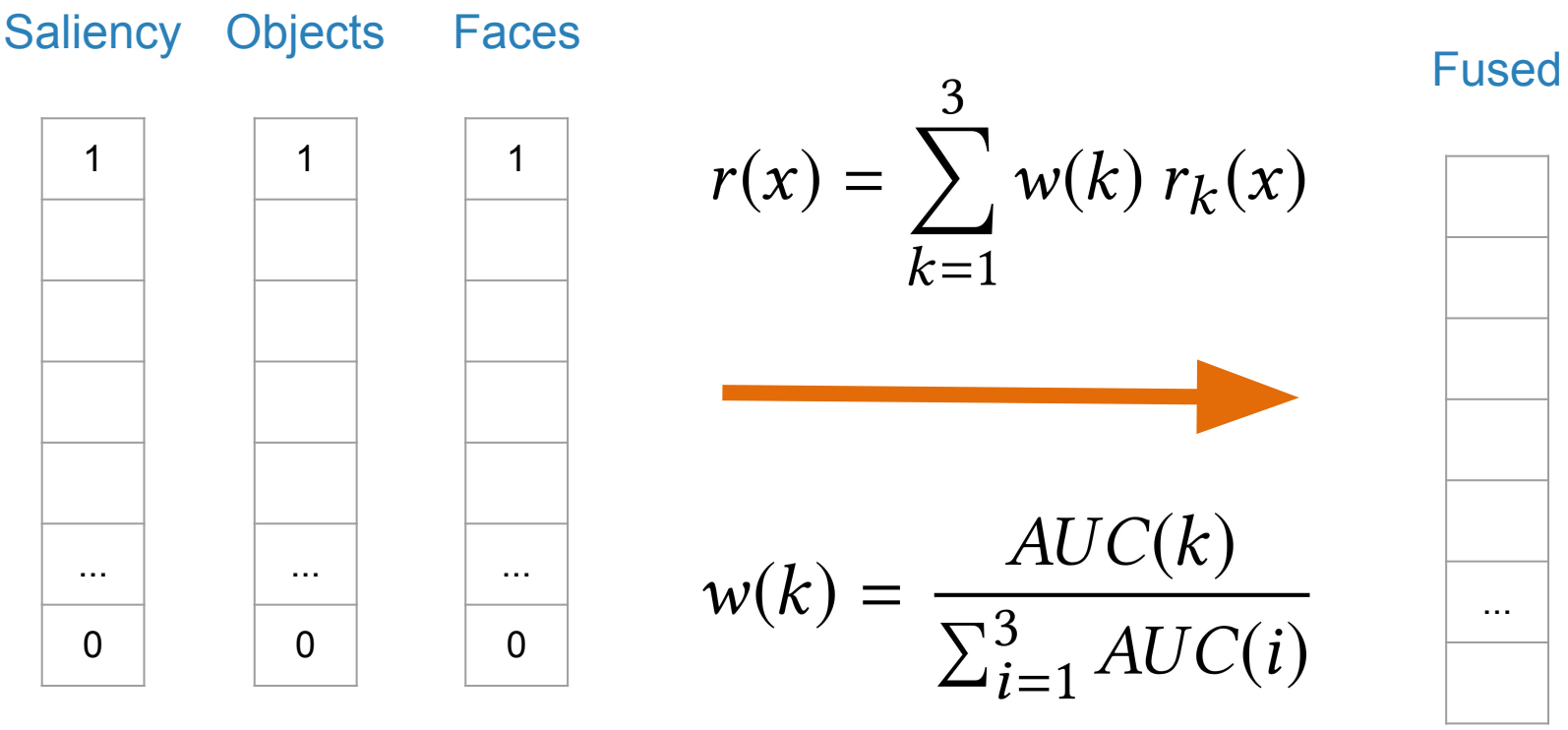
Mean averaged Sum of Max Similarities (MSMS)

Averaged across all summaries in validation set

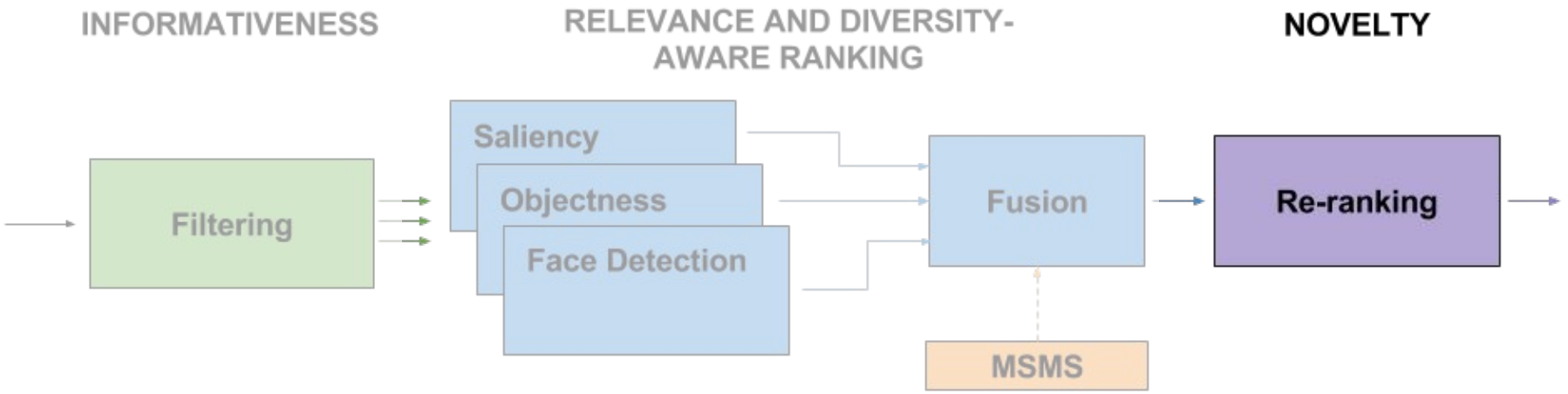


Weighted Fusion of Ranks

A weighted linear combination of scores to build r(x).



System Architecture



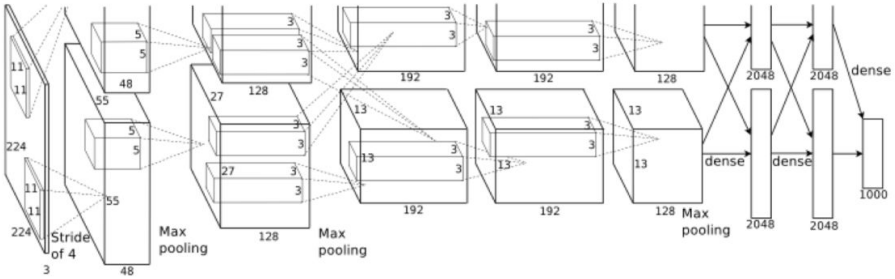
Novelty-based re-ranking

Novelty-based re-ranking

We estimate the novelty of a candidate image x^* with respect to a set Y^t based on the visual similarity to its closest match:

$$n(x^*, \mathcal{Y}^t) = 1 - s(x^*, \mathcal{Y}^t) = 1 - \max_{x_{y_j} \in \mathcal{Y}^t} s(x^*, x_{y_j})$$

Visual similarity computed as Euclidean distance between FC7 features from AlexNet-like CNN (CaffeNet).



IMAGENET

Novelty-based re-ranking

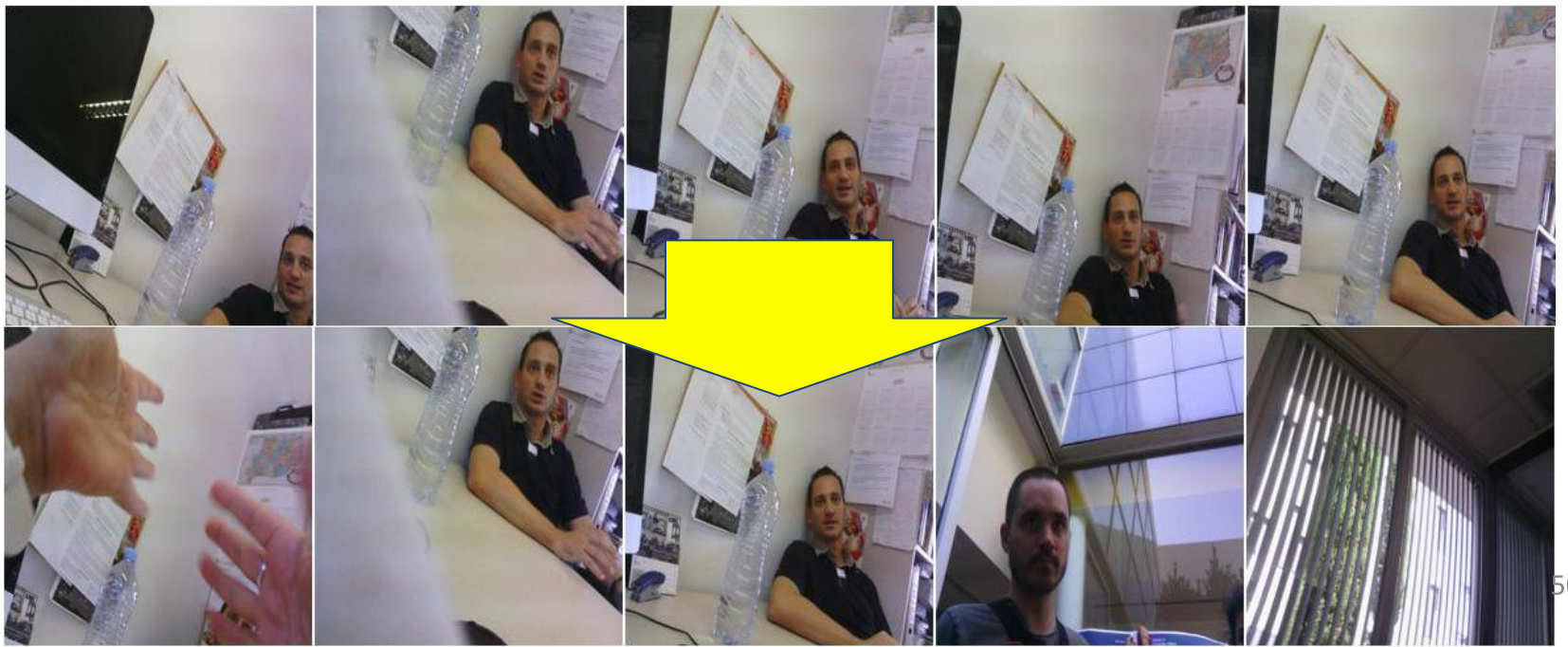
Re-ranking as a greedy selection between candidate images x^* to build a set Y of t+1 images, leveraging relevance r(x) and novelty n(x,Y).

Relevance score

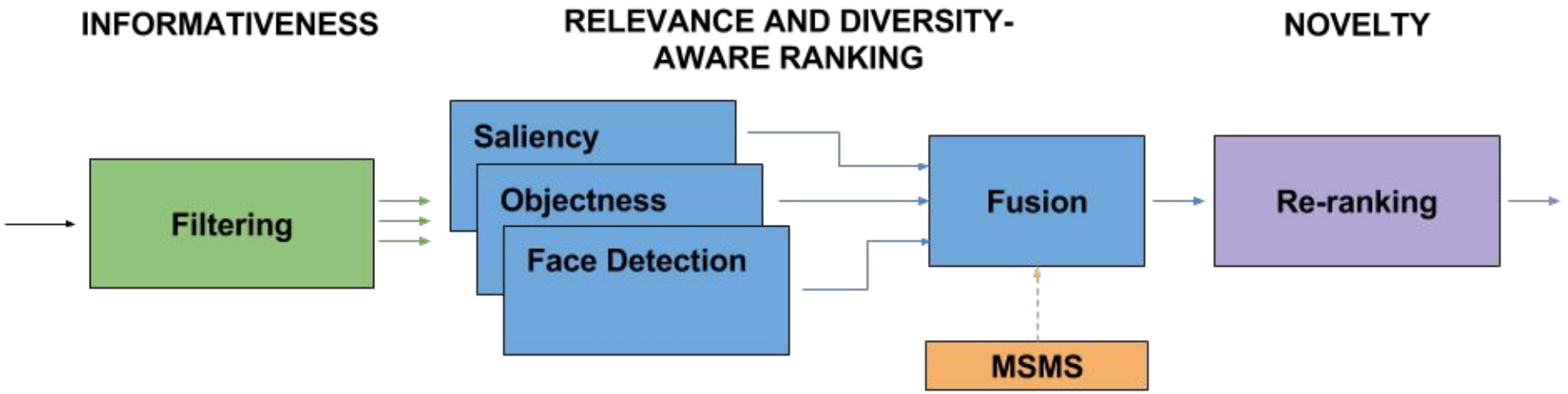
Novelty score

$$x_{y_{t+1}} = \underset{x^* \in \mathcal{X} \setminus \mathcal{Y}^t}{\arg\max} \left(r(x^*) + n(x^*, \mathcal{Y}^t) \right)$$

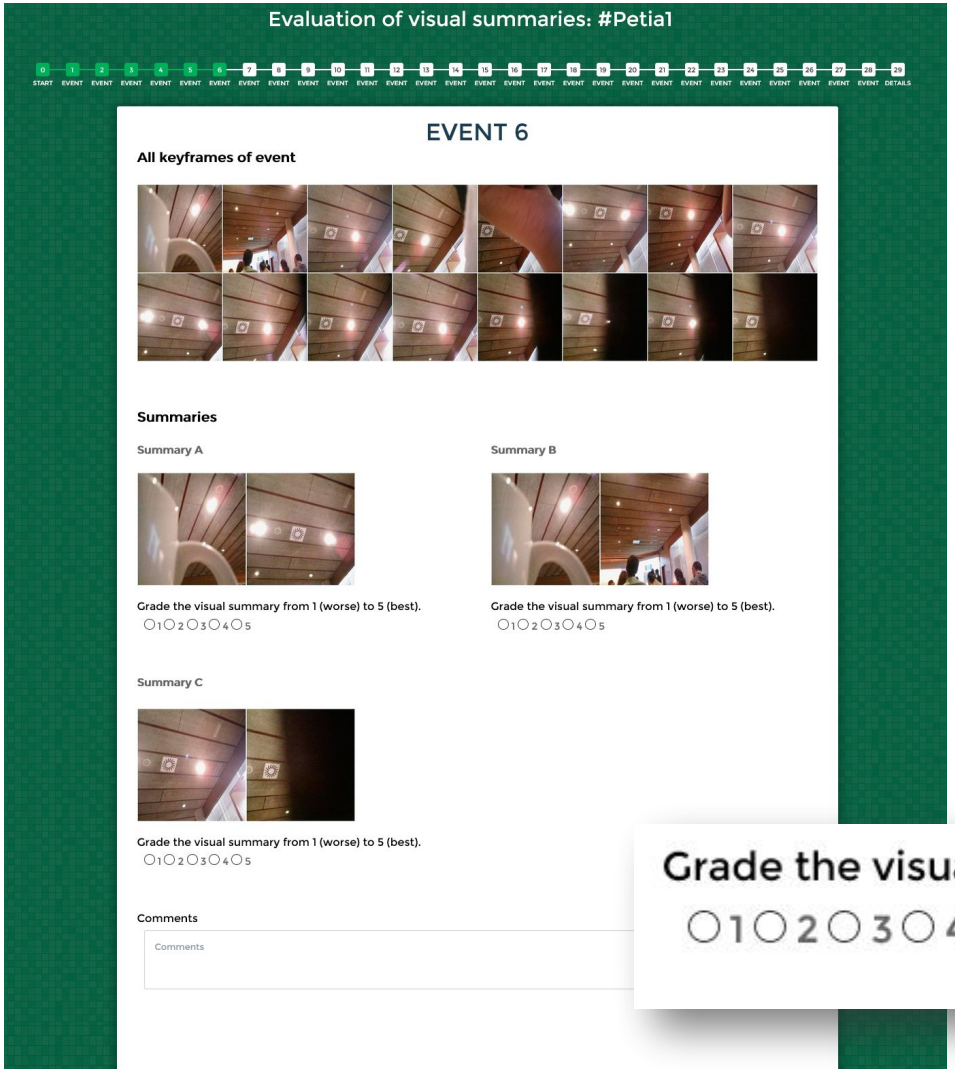
$$\mathcal{Y}^{t+1} = \mathcal{Y}^t \cup \{x_{y_{t+1}}\}$$



System Architecture



Evaluation



Feedback from medical personnel.

CST:
CONSORCI SANITARI DE TERRASSA

Grade the visual summary from 1 (worse) to 5 (best).
○1○2○3○4○5

Evaluation

Our Solution	Ground-truth	Uniform
4.57	4.94	3.99

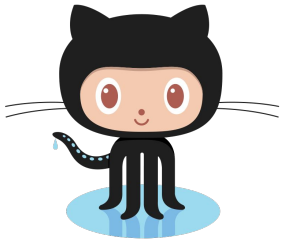
Mean Opinion Score (1 worse - 5 best)

Dataset	ImageNet	Ground-truth	Uniform Sampling
Petia1	4,36	4,93	3,39
Petia2	4,36	4,93	4,07
Marc1	4,76	4,97	4,48
Mariella	4,62	5,00	3,77
Estefania1	4,23	5,00	3,77
MAngeles1	4,92	4,75	3,92
MAngeles2	4,72	5,00	4,44
All datasets	**4,57**	**4,94**	**3,99**

Conclusions

- Ranking-based interpretation of summarization task.
- Semantic-based relevance vs Novelty (diversity).
- MSMS to assess ranked lists against unsorted sets.
- Gain in MOS validated by clinical experts.

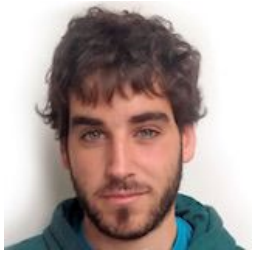
Caffe



Source code:
http://bit.ly/lta20172

DISCLAIMER: Sign included in slides as a personal decision of the speaker.

Thank you



Aniol Lidon

Marc Bolaños

Mariella Dimiccoli

Petia Radeva

Maite Garolera

Xavi Giró

UPC

Image Processing Group

Universitat de Barcelona

Barcelona Perceptual Computing Laboratory

CST

Consorci Sanitari de Terrassa

Brain, Cognition and Behaviour Group