Semantic and Diverse Summarization of Egocentric Photo Events

Aniol Lidon Baulida

Abstract

This project generates visual summaries of events depicted from egocentric photos taken with a wearable camera. These summaries are addressed to mild-dementia patients in order to exercise their memory in a daily base. The main contribution is an iterative approach that guarantees the semantic diversity of the summary and a novel soft metric to assess subjective results. Medical experts validated the proposed solution with a Mean Opinion Score of 4.6 out of of 5.0. The flexibility and quality of the solution was also tested in the 2015 Retrieving Diverse Social Images Task from the scientific international benchmark, MediaEval.

Index Terms

Egocentric vision, Diversity, Relevance, Lifelogging.

I. Introduction

A. Motivation

IN 2013, there was an estimation of 44.4 million people with dementia worldwide. This number will increase to approximately 75.6 million in 2030, and 135.5 million in 2050. In the absence of a cure for dementia, there is a real need to develop technologies to enhance the quality of life and intervention for people with dementia. Patients with dementia have difficulties remembering daily activities, names, objects, phone numbers, faces, etc. Studies have shown that using visual aids can help persons with dementia. There are several approaches in recording data about daily activities and making them available for later retrieval. Lifelogging technologies have the potential to provide memory cues for people who are in first stages of dementia. These memory cues enable the recollection of significant experiences. However, lifelogging technologies often collect a large amount of data to review (up to 3.000 per day, 70.000 per month), that makes impossible their complete observation and use. Memory cues need to be summarized.

Wearable cameras that are used to acquire egocentric images, can take 2.500-3.000 images for a day. In this set of images, some of them can appear as blurred, unfocused, dark or redundant. For example, if we meet a friend to have a coffee during one hour, during this meeting the device would capture at least 50 very similar images. Often, it is highly desirable to have tools to summarize the set of images with a few representative frames in order to optimize space fo storing and time to review.

The case scenario of this project is an application to automatically summarize a day of a lifelog (a large set of photos taken from an egocentric point of view) of a mild-dementia patient in order to remind his/her day. This project continues the work started by R. Mestre et.al. in [1]. There, the authors solved the problem of clustering sets of images from the egocentric collection of images using a event segmentation and selected the most repetitive image from an event.

Given the main goal of developing methodology for automatic egocentric summarization, our main contributions are: an iterative method that guarantees the diversity of the summary by combining different criteria: visual saliency, detected faces and detected objects, as well as a new evaluation soft metric (Mean Normalized Sum of Max Similarities), which takes into account the similarity between images to evaluate them. Note that the precision do not takes it into account. Besides the application to mild cognitive impairment patients, we tested the flexibility and quality of the system also in a different domain, by participating in the international 2015 Retrieving Diverse Social Images Task from the scientific MediaEval benchmark [2].

B. Thesis outline

Our goal is: given an event already segmented temporally from a collection of (egocentric) images, to develop an automatic method to sort the images according to their representativity. To this aim, we define a novel method to rank the list of images based on two complementary criteria: Relevance and Diversity.

Author: Aniol Lidon Baulida, aniollidon@gmail.com
Advisor 1: Xavier Giró Nieto, Image Processing Group, Universitat Politècnica de Catalunya
Advisor 2: Petia Radeva, Barcelona Perceptual Computing Lab, Universitat de Barcelona
Thesis dissertation submitted: September 2015

Fig. 1: Images acquired by a lifelogging device where objects of interest appear like: computer, mobile, coffee, hand, bicycle, person, face, flower, apple, etc.

C. Collaboration

This work is part of a multi-disciplinary project, guided by researchers from both technical and healthcare domains. The project was born from the partnership between the Image Processing Group from the Universitat Politècnica de Catalunya (GPI -UPC) and the Barcelona Perceptual Computing Laboratory from the Universitat de Barcelona (BCNPCL - UB). Mental health expertise was provided by a team of psychologists in the Grup de Recerca Cervell, Cognició i Conducta from Consorci Sanitari de Terrassa (IR3C). Physicians defined the framework for cognitive training, to determine the requirements of the image processing as well as assessed the final results of our method.

Given the technical nature of this thesis, most of the work has been developed through weekly meetings between the teams of GPI and BCNPCL, where the project was reviewed and new ideas and alternatives to the problems discussed. GPI and BCNPCL provided the off-the-shelf image processing tools and computational resources necessary to develop the project. BCNPCL and IR3C provided the annotated dataset used in this research. Meetings with IR3C allowed the definition and critical review of the project, as well as their independent assessment on the different techniques proposed in this work.

The main members that helped me to carry out the project were: Xavier Giró from GPI from UPC, Petia Raveda and Marc Bolaños from BCNPCL from UB and the group of IR3C from CST lead by Maite Garolera.

II. State of the art

A. Lifelogging

The most widely bought, known and used device, which possesses lifelogging possibilities clearly is our smartphone. However, there are also several alternative gadgets like bracelets, glasses, watches or wearable cameras, that can acquire and store information about our life during certain periods of time (hours, day, month, years). Possible lifelogging data include: our heart rate, current position, temperature, or even recorded video or images of our daily life. The most powerful cues, we can think of, obtained by a lifelogging device, are images and videos. Recently, new wearable lifelogging devices appeared such as SenseCam, Narrative, GoPro, Google Glass, Loogcie, Autographer, HP camera, etc.

These portable lifelogging devices represent the first commercial attempt to record experiences in terms of images and videos from an egocentric perspective. In fact, this trend has already been growing progressively since 1998, when Mann proposed the WearCam [3]. Then, in 2000, Mayol et al. in [4] also proposed a necklace-like lifelogging device, and in 2006 Microsoft Research started to commercialize the first egocentric lifelogging portable camera, the SenseCam [5], for research purposes.

Lifelogging devices also have more usages and more functionalities every day, and this increasing number of capabilities allows us to build more complex and useful applications. Some applications that could come to mind are: summarize the day of a person, extract nutritional information, extract physical activities, detect any kind of action that he/she performs, extract information about important objects or people in the user's environment, etc.

B. Lifelogging for dementia

Several authors [6]–[9] have studied the benefit of lifelogging cues such as egocentric images to help people with dementia to remember their memories or to help them to know about their forgotten past. Sellen et al. in [6] showed that episodic details from a visual "lifelog" can be presented to users as memory cues to assist them in remembering the details of the original experience.

To support people with dementia, Piasek et al. [8] introduced "SenseCam Therapy" as a therapeutic approach similar to the well established "Cognitive Stimulation Theraphy" [10]. Participants were asked to wear SenseCam in order to collect images of events from their everyday lives. Then, images were reviewed by the patient together with a trained therapist.

Nowadays, there are several lifelog systems, for example, the Personal Life Log system [11], which uses a combination of location sensors, physiological sensors, and real-time voice annotation to identify potentially interesting scenes in a continuous video log. However, lifelogging technologies produce large amounts of data (hundreds to thousands of images) that should be reviewed by caregivers. To be efficient, lifelogging systems need to summarize the most relevant information. Lee & Dey in [7] found that the best cues for an experience are determined by the type of the experience (such as people-based, location-based, action-based, or object-based experiences), their hybrid system allows caregivers to filter saving lots of efforts.

C. Keyframe selection on Lifelogging

Given that an event consists of many images, the challenge is to select an appropriate representative keyframe image for each one to automatically summarize the events. Doherty et al. [12] proposed a keyframe selection technique, which seeks to select the image with the highest "quality" as keyframe. First sets are split in events using temporal segmentation, then the best quality keyframe is selected using five features: contrast, color variance, global sharpness, noise, saliency and external sensors (accelerometer and light). Blighe et. al. in [13] proposed to consider three different scenarios: Static Scene, Random Scene (user walking around) and Return Scene (user walks and returns to the same point, for example, walking through the kitchen). In [14], a contribution was added taking into account neighbor frames to select the best keyframe.

D. Diversity

We are not aware of any related work within the domain of lifelogging presenting diversity-based keyframes selection in summaries, although there are pretty sophisticated approaches to diversification within text and image retrieval field. In 1998, one of the first and seminal works on diversity in information retrieval was introduced by Carbonell & Goldstain [15]. Recognizing that in the context of text retrieval and summarization, pure relevance ranking is not sufficient, the authors proposed a reranking method, called Maximal Marginal Relevance (1) that linearly combines independent measurements of relevance and diversity into a single metric, maximized in an iterative way:

$$MMR = \arg \max_{D_i \in R \setminus S} [\lambda(Sim_1(D_i, Q) - (1 - \lambda) \max_{D_j \in S} Sim_2(D_i, D_q))] \quad (1)$$

where C is a document collection (or document stream); D the document; Q is the textual query; R is the retrieved ranked list; S is the selected subset and λ is a trade-off parameter between diversity and relevance.

In image retrieval field, diversification has shown to increase user satisfaction in ranked results [16]. In [17], the authors propose a scheme for near duplicate image and video-shot detection based on color histograms. To group frames, they used Locality Sensitive Hashing, a method developed for efficiently answering approximate near neighbor queries, of the form "find all points within a given radius, R, from a query point with high probability". Leuken et al. [18] proposed different algorithms using clustering techniques to rerank.

A similar formulation to MMR of the diversification problem was given by Deselaers et al. [19] and was found to outperform a common clustering-based diversification approach, in the context of diverse image retrieval. As in MMR, diversification is achieved via optimization of a criterion that linearly combines relevance and diversity. However, [19] gives a more general formulation and uses dynamic programming algorithms to perform the optimization in addition to the greedy, iterative algorithm presented in [15].

Diversity in social image retrieval was one of the focus of the MediaEval 2013 [20] and 2014 [21] benchmarks and attracted the interest of many groups working in this field. Most participants developed diversification approaches that combined clustering with a strategy to select and return representative images from each cluster. Spyromitros et al. in [22] proposed an MMR-based approach that has the advantage of targeting the diversification problem in a more straightforward way. Compared to clustering-based approaches, the authors address a different and presumably more difficult problem. Dang-Nguyen et al. [23] contribution was to filter out non-relevant images at the beginning of the process before applying diversity.

III. Method

The goal of the system proposed in this thesis is to develop a method able to rank a set of images from a given an egocentric event, boosting to the top positions those semantically relevant images, but introducing at the same time diversity between them. The system output is a priority-based sorted list of all frames in the event. In this way, the physicians can easily choose a subset of images to show to the patient for review.

Our system is composed by three stages: **Prefiltering** uninformative images, **ranking for relevance**, and a final **diversity re-ranking**. In figure 2, we can see how the different stages of our proposal are connected.

A. Prefiltering - informativeness

Figure 3 depicts how many of the images captured from egocentric cameras have very little interest for event summarization. Some are blurred, dark or do not contain information. As Dang-Nguyen et al. [23] proposed, prefiltering useless images improves qualitatively results, because perceptually useless images are annoying. The prefiltering step aims to send to the bottom of the ranked list those images, whose quality is not informative enough.

The uninformative images are perceptually very dissimilar to the informative ones. It is very important to remove these images at this stage, because in a later one, the ones aimed for diversity (Section III-E) are going to be promoted to the first positions.

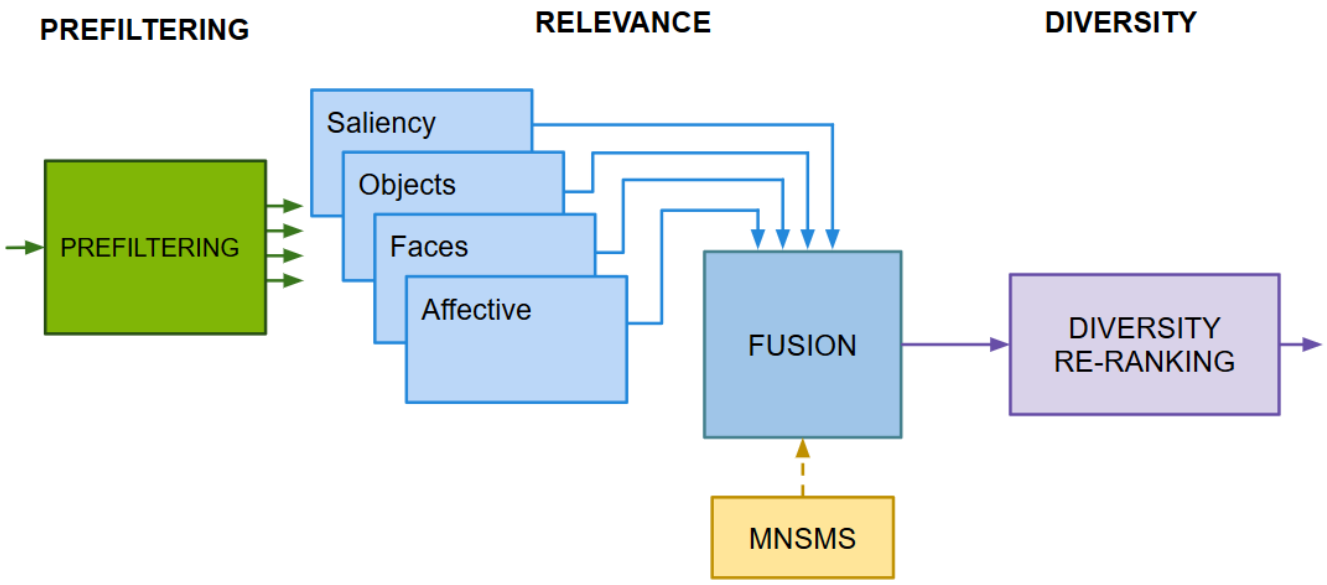

Fig. 2: Blocks diagram of the system presented in this Master thesis.

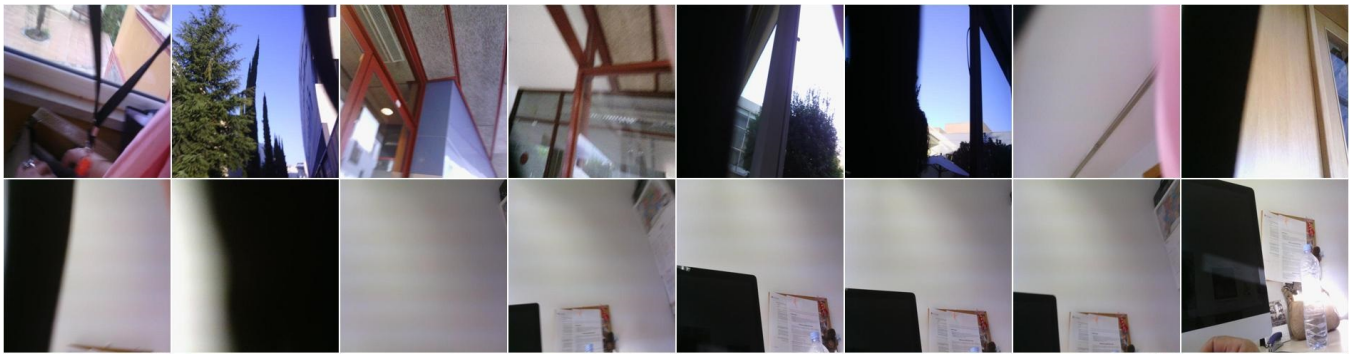

Fig. 3: First 16 frames of event 1 in Petia1 dataset. Not all images are clear, some are dark, blurred or not informative.

Images are tagged as uniformative based on a learned threshold assessed on a score. This threshold was chosen to provide a high recall, because results in this stage are irreversible in the pipeline.

We have tested two approaches for prefiltering informativeness images:

- Our first approach is based on hand-crafted estimators of darkness, blurriness or burned images. To estimate if a image is blurred, we consider the discrimination between different levels of blur perceptible on the same picture using the algorithm proposed by Crete et al. [24]. In order to estimate the darkness or brightness, we compute the mean of the image by averaging all pixels. Thresholds for blurriness, darkness and brightness are obtained training from manual annotations for each.

- Our second approach, which gives better results, is to train through deep learning an **Informativeness Network** based on HybridNet [25], a CNN trained with objects from the ImageNet dataset [26] and locations from the Places dataset [25]. HybridNet is fine-tuned in two classes: *relevant* and *irrelevant*, as labeled by human annotators. Final uninformative selection is only from images that the network is completely sure, the score for uninformative class is one. The network used in this second approach was trained by Marc Bolaños from the BCNPCL group [2] and publicly available online[1].

B. Ranking for relevance

The second step of our system is to rank the informative frames using estimators of relevance. The first question that appears on our mind is: what we consider as **relevant**. The team of psychologists described the set of relevant keyframes in the event (event summary) considering relevant frames in a summary as:

[1]https://imatge.upc.edu/web/resources/iterative-reranking-relevant-images-software

- Non-repeated images.
- Unusual images.
- Image representative of an activity. (*What?*)
- Image containing social interactions. (*Who?*)
- Images containing information from the environment. (*Where?*)
- Images that contain information on when an event has occurred. (*When?*)
- Images that tell how activity occurred. (*How?*)

Our work in relevance only focuses on frame level selecting relevant frames. Due to time and scope of this project only two criteria are inquired: frames that answer the question *What?* and Who?. We searched off-the-shelf algorithms that solved these questions and the ranking obtained with each of them was later fused.

1) Semantic relevance: In egocentric images, activities can be described by the objects manipulated by the wearer in the scene and the spatial information. For representativity of activities, answering question *What?*, we use *Saliency maps* for detecting the amount of salient zones and *Object candidates / Object detector* for estimating the amount of objects in the image. Social interactions, related to question *Who?* are described by other humans. We use *Face detection* to detect presence of other humans and *sentiment analysis* to detect emotions in images. Applying these semantic analysis modules in each frame, we obtain a score, which we will use for ranking.

A **Saliency map** is a model that explains eye fixations on a visual scene. The prediction of saliency areas in images has been traditionally addressed with hand-crafted features based on neuroscience principles. To obtain a relevance metric for saliency maps, we make the following assumption: *As more interesting zones an image has, more relevant it is.* Saliency maps systems output is a grayscale image, applying the mean of all pixels, we can obtain a score for each frame related to the amount of relevant zones. We have used SalNet, a deep learning network upon CaffeNet architecture, provided by Kevin McGuinness from Dublin City University (to be published). As another approach for saliency estimation, we have tried to focus on the relevant zones at the center of the image multiplying a Gaussian centered to the saliency pixels. The idea is to give more weight to the pixels in the center.

The goal of using **objects** for relevance is based on the assumption that the more objects in an image, the more relevant the image is. In addition, as more defined is an object, higher will be the confidence that a computer vision algorithm provides. We have tested the two types of algorithms aimed at objects:

- **Object candidates** algorithm generates a list of object proposals. Each proposal includes the confidence to be an object. The sum of confidences of all proposals is proportional to the definition of objects and the amount of objects found, so this sum is the score for ranking. We expect to have all object proposals surrounding real objects. We have used the state-of-the-art **Multiscale Combinatorial Grouping (MCG)** algorithm [27].
- **Object detector** algorithm detects, which objects are in the scene and gives a confidence of each of them. We have used the **Large Scale Detection through Adaptation (LSDA)** algorithm [28] for object detection and recognition. Section III-D describes in detail the usage of this algorithm. LSDA algorithm has a No-Maxima-Suppression (NMS) step, which leaves only a few objects candidates, clearly containing an object. With the sum of all confidences from the objects detected, we obtain the score for ranking. In contrast to using the LSDA as a similarity measure, we want to keep both NMS steps in order to detect only clear objects, and sum confidences of these objects.

Face Detection algorithms was used to estimate social interactions. We used an off-the-shelf face detector by Zhu et al. [29], which provides a confidence of the detected faces. To obtain a score for each frames, we summed the exponential of confidences obtained, because confidences are not always positive. When multiple face were detected, the score of that frame increased due to the exponential sum.

Doctors consider that images, where a sentiment is expressed, are also very important for memorability. The basic idea of this affirmation was that our mind remembers situations in function of the sentiment felt. In Computer vision, there are several studies of **affectivity and sentiment** analysis of an image. The emergence of big databases has allowed Convolutional Neural Networks to learn emotions. We tested a CNN [30], which fine-tunned CaffeNet with a dataset for sentiment prediction, which has two classes (positive an negative) labeled by 5 human annotators. To rank each frame, we tested three ways to obtain the score:

- Considering the highest positive image as relevant.
- Considering the highest negative image as relevant.
- Considering the extreme sentiments as relevant following the equation 2, $positive$ and $negative$ classes are complementary and their sum is one.

$$Score = 2 \cdot |positive - 0.5| \tag{2}$$

C. Late fusion of relevance ranked lists

The ranked lists obtained for each relevance method were fused in a single one. Each method does not have to had the same influence in the final relevance, because sometimes cues can contradict each other or some can give more information. It is necessary to introduce weighting for each method between cue fusion. These weights were learned using MNSMS metric, which basically estimates the accuracy of each relevance method when used separately. MNSMS is explained in detail in Section IV-A.

For obtaining final relevance of an event, we need to follow four steps:

1) Estimate each SOFA relevance (Saliency, Objectness, Faces, Affective) to all the frames our given event.
2) Obtain a ranked list for each method sorting the scores obtained.
3) Apply a normalization strategy, based on ranked or on scores. It is required to truncate the initial lists to the top N results and normalize them. In this thesis, two options are considered:
   - Normalization by rank:
   $$\overline{R}_k(n) = \frac{N + 1 - R_k(n)}{N} \tag{3}$$
   - Normalization by score:
   $$\overline{R}_k(n) = \frac{R_k(n) - \min(R_k)}{\max(R_k) - \min(R_k)} \tag{4}$$
   where R_k is the method result list (rank or score).
4) Aggregate the normalized scores ($\overline{R}_k$) weighting them (w_r) to generate a single ranked list.
$$R_s(n) = \sum_k (\overline{R}_k(n) \cdot w_r(k)) \tag{5}$$

Weights are learned using MNSMS metric, comparing the performance of each relevance method. Each method is evaluated separately using the MNSMS metric (see section IV-A). From the obtained **Area Under the Curve (AUC)** of each method m, we obtain each weight $w(m)$ by:

$$w(m) = \frac{AUC(m)}{\sum_{k=1}^{M} AUC(k)} \tag{6}$$

D. Similarity

The obtained ranked lists will be later re-arranged based on an algorithm that requires computing the similarity between images. Similarity is an abstract concept, so we can consider two images being similar depending on the adopted criteria.

In this thesis, we have considered three similarity measures. Two of them are perceptual similarities and the third one is a semantic one. When we talk about **semantic similarity**, we mean that the feature vector describes an abstraction in the image in terms of concepts. **Perceptual similarity,** instead, is more based on a low-level features, which describe an image in terms of their appearance. For example, in Figure 4, we have four images, semantically all images are equal (chairs), but perceptually, we have two types of chairs.

All distances are obtained applying the Euclidean distance (norm L_2) between two feature vectors from a deep learning network. Distances are normalized dividing by the maximum of all crossed Euclidean distances in the dataset. To move from distances to similarities, we need to subtract them to one.

The Convolutional Neural Network architecture employed in all our similarities is CaffeNet, a slight modification of the ILSVRC 2012 winning architecture [31]. This network, which was originally designed and trained for the task of object recognition, is composed by 5 convolutional layers and 3 fully connected layers. The two first convolutional layers are followed by pooling and normalization layers, while a pooling layer is placed between the last convolutional layer and the first fully connected one. The experiments were performed using Caffe, [32] a publicly available deep learning framework.

Fig. 4: Example of perceptual and semantic similarity.

Our first perceptual similarity measure is trained with **ImageNet** [33] database using CaffeNet Architecture. ImageNet contains 1000 different objects organized according to the WordNet hierarchy. Layer 8 has been removed to keep a higher lever of extraction, removing last layer objects are described in a perceptual way. The reason we have chosen ImageNet database is, because we expect a correct description of a picture describing the objects appearing.

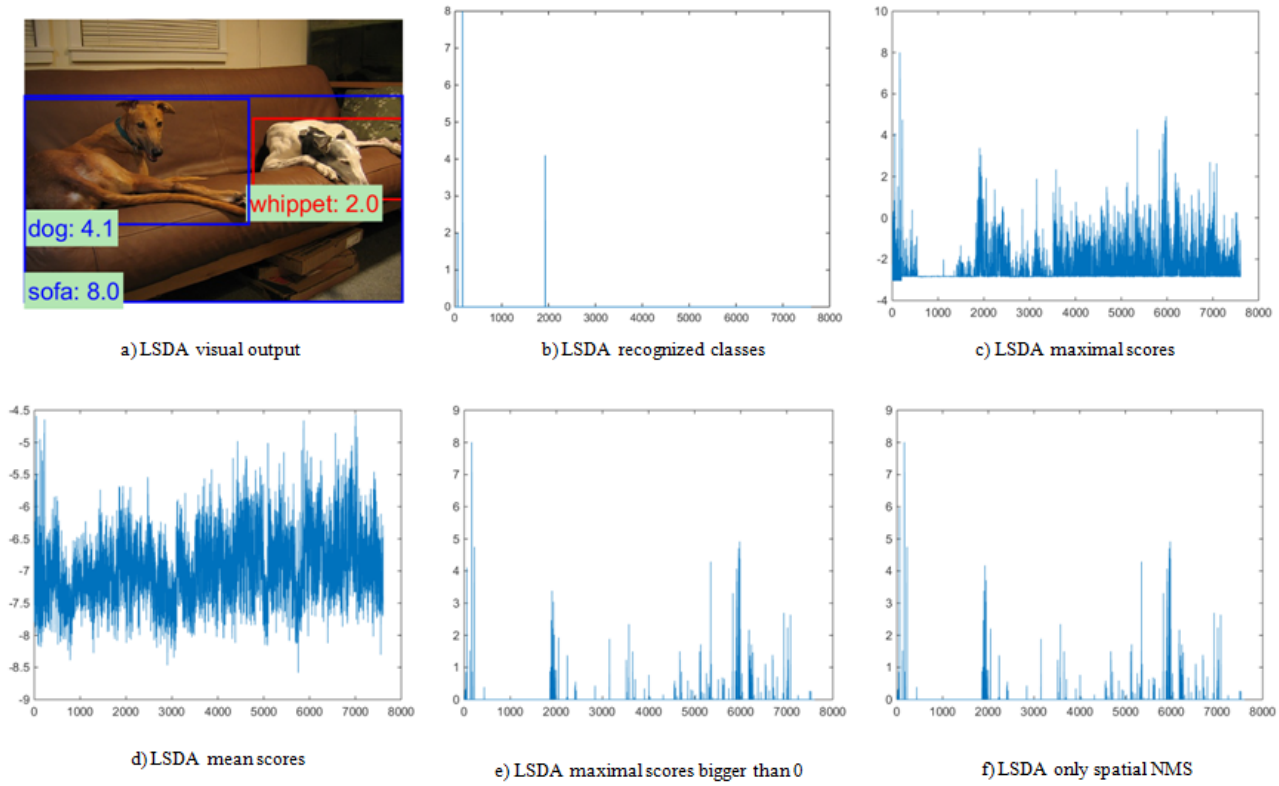

Fig. 5: LSDA Feature vector of the image a applying different merging methods of the scores obtained in bounding boxes.

Our second perceptual similarity measure is trained on **Places** database [25], which also uses CaffeNet architecture. Places database contains 476 place categories. Layer 8 has also been removed to work in a perceptual space. We expect that Places network describes the pictures in a scene way, so it may be interesting, when we are focusing on egocentric point of view.

Our semantic similarity measure is based on **Large Scale Detection through Adaptation**. LSDA algorithm increases the number of classes to 7600. The algorithm is able to recognize each class at the same time it is able to detect it. To do so, researchers from Berkley were able to create an algorithm, which learns the difference between the two tasks and transfers this knowledge to classifiers for categories without bounding box annotated data, turning them into detectors. The potential of this method was to enable detection for the tens of thousands of categories that lack bounding box annotations, although they have plenty of classification data.

LSDA algorithm passes all bounding boxes detected by Selective Search [34] algorithm to the convolutional neural network. After recognizing objects in each bounding box, two post-processing steps of NMS are needed to remove overlapping bounding boxes (spatial NMS) and to assign to each bonding-box left to just one class (semantic NMS). For adapting LSDA to feed our needs, we made several configurations:

- *Using LSDA as the original implementation, constructing the feature vector with detected classes.* In figure 5, we can see that the obtained feature vector (b) obtained is very poor, because we only keep information on the final detection. If, for example, we have two similar images, but in one image an object is not correctly detected, the Euclidean distance between both images is going to be high. More information needs to be kept for our application.
- *Avoiding both NMS steps and taking the maximal scores for each class in all the boxes found given by selective search algorithm.* LSDA gives for each bounding box a vector of scores for each category. As our application does not need to know, where the objects are, we can just take the higher scores for each class. As we can see in figure 5c, nearly all classes have a score now, so we have increased information.
- *Avoiding both NMS steps and taking the mean of the scores for each class.* Instead of taking the maximum, we have tried to compute the mean score of all bounding boxes. As we can see in figure 5d, it does not seem to be a good idea, because a lot of noise is introduced into the signal.
- *Avoiding both NMS steps and taking the maximal positive scores for each class.* Trying to improve maximal scores solution, we found out that considering negative scores was not a good idea, because scores under 0 mean that the object is not found. As we can see in figure 5e, we filtered out the majority of noise. The problem, we have found in this proposal, is that the information about the amount of objects repeated in the scene is lost.

Fig. 6: Eight first frames of event 9 in the "Mariella" dataset. There is a high information redundancy between frames.

- *Using only the spatial NMS to remove overlapping bounding boxes, but maintaining all detentions in these bounding boxes.* Our last approach is to maintain the NMS in overlapping bounding-boxes to keep windows only around clear objects, filter out scores below 0, which are not considered objects for their low score and sum the remaining scores in all bounding boxes. This approach allow us to increase scores on repeated objects, remove confusing objects and at the same time avoid to label each bounding box to a single object.

To decide which of the previous proposals was the best, we tried the different configurations in a closed scenario, which we will explain later on section IV.

E. Diversity re-ranking

Doctors described a summary in a semantic level as a set of images providing meaning. A summary provides meaning by:

- Being representative of the action.
- Being diverse.
- Containing important information.
- Containing interesting (salient) information.
- Describing a story. Brain needs narrative. If brain does not have it, it will invent it. People do not remember just objects, people remember narratives between them.

We introduce diversity in order to add meaning to a summary, it helps to tell the story removing unnecessary information.

Diversity re-ranking stage is vital to remove redundancy between images. We can see in figure 6, a subset of frames labeled by doctors as *"working in computer"*. A huge amount of redundancy appears inside egocentric events. Although in nearly all frames, we can clearly see the action (we consider them as relevant), redundancy needs to be decreased.

Diversity re-ranking requires the similarity comparison of all images in the event. In section III-D, we have discussed the three similarity measures used in this thesis. As similarity measure, we propose three different distances based on Deep learning: a similarity based on layer 7 of ImageNet, a similarity based on layer 7 of Places (both distances perceptual) and a third similarity measure based on LSDA objects [28] as a semantic distance.

Our proposal for diversity re-ranking is to avoid using clusters and use a soft reranking as proposed by Carbonell & Goldstain [15]. Our goal for this diversity stage is to reduce neighboring frames intra-similarity respecting relevance obtained in the previous stages.

Taking as reference the work from Spyromitros et al. [22], who proposed an adaptation of the query-based MMR (1), we implemented an iterative 5 steps algorithm to apply diversity as a similarity measure. We called it **Re-ranking by Soft Max Diversity**. This algorithm starts with an empty list D and relevance ranked list R. The steps are the following:

1) The first element of R goes to the bottom of list D and it is removed from the set R.
2) A score is given to each frame of R using a normalization by rank (3), where the first element has a score of 1 and last is 0.
3) The scores of R are updated subtracting the similarity between each frame and the most similar frame of the set, D.

$$\overline{R}_s(k) = R_s(k) - \max_{n=1}^{N} sim(R(k), D(n)) \tag{7}$$

4) R is re-ranked with updated scores.
5) Iteratively, all steps are repeated, while there are remaining elements in R.

The D ranked list has the final ranking.

The drawback of the Re-ranking by soft diversity is that we only can consider one similarity metric. The similarity between two images can be given by a lots of factors. In Figure 7, we can say that all images are similar in terms of pose, shot, person and type of dress, but images are dissimilar in terms of dress color or background. With this example, we want to illustrate that sometimes we need to take into account more than one similarity measure.

Fig. 7: Angela Merkel chancellor of Germany with several dresses.

As a new contribution from this work, we propose a novel strategy to fuse diversity measures using more than one similarity metric. **Re-ranking by Soft Max Diversity Fusion** is the evolution of RSMF algorithm and uses the maximum similarity measure, when updating scores. As in the relevance algorithm, a weight for each similarity measure will be needed. S is the number of similarity measures, we want to apply and w_s, the weight for each similarity. The steps, the algorithm follows, are:

1) The first element of R goes to the bottom of list D and it is removed from set, R.
2) A score is given to each frame of R using a normalization by rank (3).
3) The scores of R are updated subtracting the weighted sum of similarity between each frame and the most similar frame of set, D.

$$\overline{R}_s(k) = R_s(k) - \sum_{s=1}^{S} w_s(s) \cdot \max_{n=1}^{N} sim_s(R(k), D(n)) \tag{8}$$

4) R is re-ranked with updated scores.
5) Iteratively, all steps are repeated, meanwhile there are remaining elements in R.

The ranked list, D has the final ranking.

IV. Experiments

Experts' manual evaluation is the best evaluation system, because they exactly know how the *Cognitive Stimulation Theraphy* [10] works and which is the system that fits their therapy. Doctors are considered as experts evaluators. The team of psychologists, which evaluated our results, is from the IR3C (Grup de Recerca Cervell, Cognició i Conducta) group of CST. Section IV-E presents the evaluation questionnaires based on Grauman et al. [35] evaluation and later used by R. Mestre et. al. [1].

Although experts' evaluation is the optimal one, it presents the substantial drawback that they require a great amount of human resources which, in the case of medical experts, is specially expensive, difficult and subjective. Intermediate automatic experiments (section IV-C) were used for system development using the Mean Normalized Sum of Max Similarities (MNSMS), which is introduced in Section IV-A.

A. Evaluation Metrics

The first problem, we found, when raising the project was, how we were going **to evaluate the results**. We observed that the goal to decide the best summary was not always easy. Different people have different opinions in which to order the set of images that summarizes an event. So we were treating a very subjective problem. To solve this ambiguous problem, we asked physicians to manually build visual summaries from certain event, which were already segmented. We assumed that the frames selected by the doctors would be the *relevant* for our problem and, noticeably, that the non-selected would be *non relevant*.

This ranked list was automatically assessed based on the following metrics:

1) Precision at K (P@k):: Estimates the proportion of relevant images contained in a selection of the first k elements in the ranked list. (Equation 9)

$$P(k) = \frac{|\{relevant\} \cap \{retrieved\}|}{k} \tag{9}$$

2) Average Precision (AP): The precision at those positions containing the relevant images is summed and divided by the total amount of relevant images in the dataset. This measure approximates the area under the Precision-Recall curve of the ranked list. (Equation 10)

$$AP = \frac{1}{N} \sum_{k=1}^{N} P(k) \cdot rel(k)), \qquad rel(k) = \begin{Bmatrix} rel(k) = 1 \Leftrightarrow k \in R \\ rel(k) = 0 \Leftrightarrow k \notin R \end{Bmatrix} \tag{10}$$

Fig. 8: Subset of images from the "Petia1" dataset showing the similarity between images.

3) Mean Average Precision (MAP): The APs obtained in the M events contained in the dataset are averaged to compute a final metric that measures the overall goodness of the solution (Equation 11).

$$MAP = \frac{1}{M}\sum_{i=1}^{M} AP(i) \tag{11}$$

Precision is a good method to obtain an objective measure, but as we said before, our solution is subjective. For example, in Figure 8, we have four images, where we have selected two images, both give nearly the same information to the patient. In the case that psychologists choose the green one, but our algorithm selects the blue, MAP will fall although our result is equally acceptable.

For this reason, we decided to propose a new metric, we call **Mean Normalized Sum of Max Similarities (MNSMS)**, which takes as reference the annotations from the experts and penalizes in function of the similarity to the ground-truth. Now returning to the example in figure 8, in the case that the blue image is chosen instead of the green one, MNSMS value will be high, but if the first image (coffee machine) is chosen, MNSMS value will fall down. To understand the final metric in detail, we will explain the partial metrics *Sum of Max Similarities*, and the *Normalized Sum of Max Similarities*.

4) Sum of Max Similarities (SMS): First, we focus inside an event. We have a sorted list of frames (the obtained result, we want to evaluate), that has to match to a list of ground-truth keyframes. Taking the first n frames and matching each ground-truth keyframe to the most similar frame of this cropped list, we can compute the similarity committed if we only keep n frames. Summing the similarity at each step n, we obtain a curve that increases in n. This is due to the fact that the more frames we get, the more is the probability of matching in all ground-truth keyframes.

To do the similarity matching, first, we need to compute the similarity matrix between the ground-truth keyframes and all the frames. The computation of the similarity between images is discussed in Section IV-A. The similarity matrix will present a value close to zero between two different images, and close to 1 between two similar ones. So, as we have the ground-truth frames in our list, ones will appear in the similarity matrix.

To perform the similarity matching, we proposed two approaches:

- Hungarian Algorithm
- Max Similarity

Our first approach is to enforce each ground-truth keyframe to match a different frame. To do so, we use a well established **the Hungarian method** [36]. The Hungarian method is a combinatorial optimization algorithm that solves the assignment problem in polynomial time. Using this algorithm, we enforce diversity at evaluation, but it poses new challenges. The computational time to apply the method in large events is high; and we can only perform steps (n) bigger than the number of ground-truth images. For reducing the computational time of Hungarian algorithm, we used a dynamic fast implementation [37] instead of the classical one.

The second problem is produced in the first steps ($n < num(gt)$), because Hungarian is avoiding to repeat images, but there are not enough elements. We solve this shortcoming by two different approaches: Our first approach is to consider that the matches that are not found, correspond to similarity, 0. The second approach is to ignore the curve below $num(gt)$. Both approaches have their drawbacks. Considering not-found matches as similarity zero produces big jumps in first steps. And ignoring the curve below $num(gt)$ does not consider how the curve works in the beginning, where the most relevant information should be. Also, when normalizing the curves to merge all events (explained in the next section), both approaches are not satisfactory, because they depend too much on the number of ground-truth frames.

The second approach in the similarity matching does not enforce to match different frames. However, it takes the assumption that the ground-truth is already diverse. The **Max Similarity** method works as follows: each ground-truth frame matches to

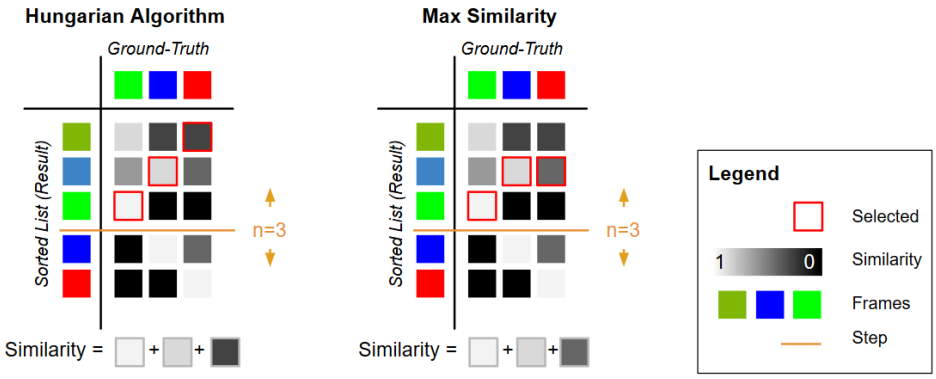

Fig. 9: Visual representation of the Hungarian Algorithm and the Max Similarity, at step $n = 3$.

the most similar (maximum similarity) frame in the cropped list (with n first frames). The similarity at n is the sum of the most similar frame of each ground-truth keyframe. If we consider a diverse ground-truth repetition will only occur, when the step n is low. Max Similarity will be used in this project instead of Hungarian algorithm, because it is computationally faster and it does not produce errors in the normalization process.

In Figure 9, we can see a visual representation of both proposals at step $n = 3$. Each frame is represented by a color and similarity is represented by a gray color. Both algorithms select the most similar frame to the ground-truth. The Hungarian algorithm has the constraint that a frame only can be selected once. If we move the step to $n = 4$, we see that the sum of similarities decreases, because the blue frame will be selected.

5) Normalized Sum of Max Similarities (NSMS): Moving to a dataset level, we need to join the SMS curves from all events. As we noticed, all the curves need to be normalized in both axes before averaging them. The normalization in the SMS axes is easy, we just need to divide by the number of ground-truth frames to put the maximum to 1. To normalize in the step axes, we will have to reshape the number of samples. Our first approach is to re-sample the curves using interpolation and decimation with an intermediate anti-aliasing filter. We observe that this produces fake shapes of the curve that affects the results. Our second approach is to assign at each new sample the value of the nearest old sample at ceiled proportional position.

6) Mean Normalized Sum of Max Similarities (MNSMS): Once we have normalized each event curve, we can apply the mean of all events to obtain a curve representing the behavior of the method in the dataset. To compare two different methods, we use the Area Under the Curve. The AUC is normalized dividing to the number of samples chosen in the normalization step to present it between 0 and 1. The bigger the AUC is, the better the method is. The ideal method is the one, which has an AUC of 1.

7) Physiologists feedback: Final results are evaluated presenting a questionnaire to the IR3C group. The questionnaire presented to the psychologists, is explained in detail in section IV-E. Our final evaluation approached is based on a previous knowledge of the amount of images necessary to summarize an event, based on the doctors annotations. When presenting results to doctors, the final ranked list of images is truncated at the amount of images estimated as necessary for the physicians. This truncation is necessary to convert lists to summaries.

B. Database

Images used in this thesis come from seven datasets corresponding to seven days recorded by a *Narrative Clip* camera (see figure 10). This camera is small and light, and can be easily worn on the clothes. It takes an image, by default, every 30 second.

Collections contained in the dataset collected by BCNPCL are detailed in table I, structured according to the person wearing the camera. In the table, one can find the number of events for each set and the number of images of each. ”MAngeles1” and ”MAngeles2” sets are mostly outdoor, meanwhile other sets are in indoor scenes.

Fig. 10: Narrative Clip camera.

C. Intermediate experiments

Intermediate experiments using NMSMS allow us to make decision for the final configurations evaluated by the psychologists. In all experiments, unless it is not specified, similarity measure used for NMSMS is obtained from the Euclidean distance

Dataset name	Number of events	Number of images
"Petia1"	27	1388
"Petia2"	13	684
"Marc1"	29	885
"Mariella"	12	586
"Estefania1"	12	1388
"MAngeles1"	12	428
"MAngeles2"	18	609

TABLE I: BCNPCL datasets description.

between the CaffeNet features. We chose the CaffeNet option, because the features
are widely extended in the Computer vision field and several studies [38] have shown its good performance for extracting multi-purpose visual features.

In *Method* section (III), we proposed several approaches for our problems. The goal of this section is to decide, which approach works better for our application.

In the **prefiltering** stage III-A, we proposed two approaches: a *basic thresholding* on dark, blurred or burned images and a deep learning approach training an *Informativeness Network* with manual annotations of non-informative images. The three experiments, we are going to evaluate regarding the prefiltering stage, are:

1) Directly evaluate the event, the ranking of images is just temporal (prefiltering, relevance and diversity stages are switched off).
2) Prefilter events using the *basic thresholding* and evaluate results (relevance and diversity stages are switched off).
3) Prefilter events using the *Informativeness Network* and evaluate results (relevance and diversity stages are switched off).

For **saliency relevance method**, we proposed to multiply saliency maps by a *Gaussian centered* on the image. Two experiments are proposed to solve this problem:

1) Rank events using saliency relevance and evaluate results (prefiltering and diversity stages are switched off).
2) Rank events using saliency multiplied by a centered Gaussian and evaluate results (prefiltering and diversity stages are switched off).

For the **object relevance method** , we need to determine which proposed algorithm gives better performance: MCG as an object candidates, and LSDA as an object detector. Prefiltering and diversity stages are switched off. The experiments proposed to solve this problem are the following:

1) Rank events using objects' MCG relevance and evaluate results.
2) Rank events using objects' LSDA relevance and evaluate results.

For the **affective relevance method**, we have to determine, which sentiment is considered relevant: positive, negative, or extreme. Similarly to the other relevance methods, for these experiments prefiltering and diversity stages are switched off. The experiments proposed to consider this problem, are the following:

1) Rank events using affective positive relevance and evaluate results.
2) Rank events using affective negative relevance and evaluate results.
3) Rank events using affective extreme sentiment relevance and evaluate results.
4) Rank events using random relevance and evaluate results.

In the **relevance fusion**, it is necessary to determine, which normalization is better. Similarly to the other relevance methods for these experiments, prefiltering and diversity stages are switched off, relevance uses Saliency without centered Gaussian fused with Faces. The weighting for each method is the same, 0.5. The experiments differ in the normalization method:

1) Normalization by rank.
2) Normalization by score.

In the case of **LSDA used as similarity** to assess the diversity, it is necessary to determine, which configuration has better performance. LSDA configurations were described in sectionIII-D. Using MNSMS metric is a *poisoned* solution, because we are evaluating a similarity measure using ImageNet similarity measure, so solution obtained possibly is the one that fits better in ImageNet. For these reason in this experiment, we are taking a look also to the Mean Average Precision Metric to ensure to not commit a big mistake. For these experiments, prefiltering and relevance stages are switched off. The experiments proposed are:

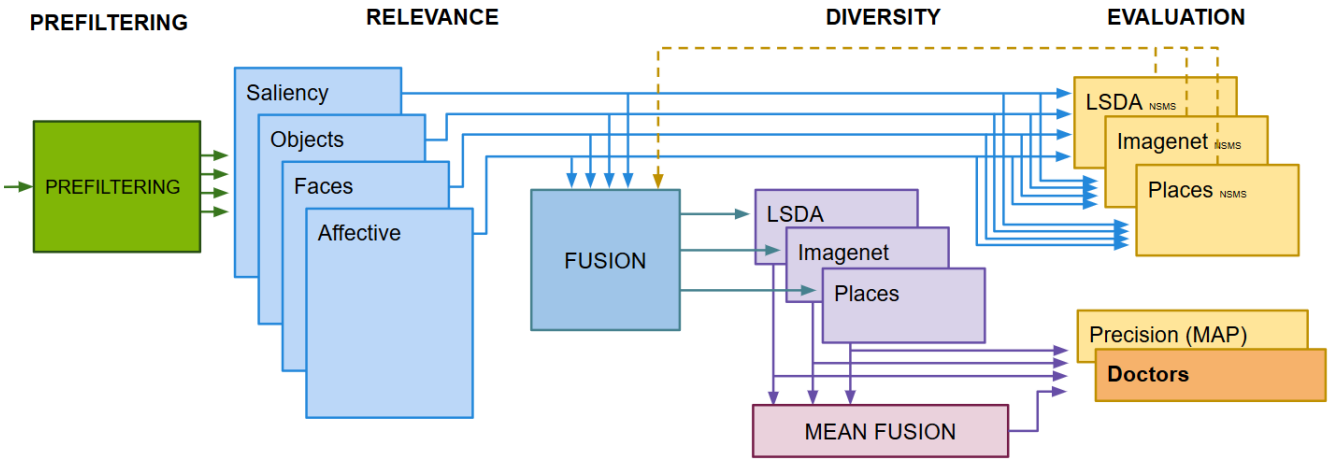

Fig. 11: Pipeline of experiments

1) Directly re-rank events using soft max diversity with original LSDA (two NMS steps) as similarity metric.
2) Directly re-rank events using soft max diversity with maximum score LSDA (without NMS) as similarity metric.
3) Directly re-rank events using soft max diversity with mean score LSDA (without NMS) as similarity metric.
4) Directly re-rank events using soft max diversity with maximum positive score LSDA (without NMS) as similarity metric.
5) Directly re-rank events using soft max diversity with spatial NMS only LSDA as similarity metric.

D. Final configurations

Final configurations, that has been presented to the doctors, can bee seen in figure 11. Four final configurations are presented.

In figure 11, we can see the three stages of our system, prefiltering, relevance and diversity. All final configurations start with a common prefiltering stage (in green), SOFA relevance rankings are computed (in blue) and are fused (in dark blue) with learned weights (in yellow) using MNSMS's AUC. Each configuration uses the same similarity metric, which has been trained to diversify (in purple). Different configurations are:

1) A first prefiltering stage, SOFA relevance fused with learned weights in LSDA MNSMS's AUC and diversified with LSDA similarity.
2) A first prefiltering stage, SOFA relevance fused with learned weights in ImageNet MNSMS's AUC and diversified with ImageNet similarity.
3) A first prefiltering stage, SOFA relevance fused with learned weights in Places MNSMS's AUC and diversified with Places similarity.
4) A first prefiltering stage, SOFA relevance fused with equal weights and diversified with Re-ranking by Soft Max Diversity Fusion algorithm with equal weights.

E. Expert Evaluation

The four configurations explained in section IV-D were evaluated manually by experts. We made two rounds of questionnaires. In the first round, we asked the doctors to compare the four configurations and in the second round we asked to compare the best configuration to their own ground-truth and to an uniform sampling summary.

One can note that our system does not return a summary, it returns a sorted list of frames, where relevance and diversity are condensed in the first frames. To convert the system output to a summary, a crop of the list is needed. Each event needs a different number of images. However, developing a method for automatic determination of the optimal length of summaries is out of the scope of this thesis. In the future work section (VI-B), we will discuss some ideas how to approach it.

Presented summaries are cropped with the length of the given ground-truth. Our brain needs to find the story between images, for this reason, it is very important to present frames temporally. So, each summary is built with temporally sorting of the N (length of ground truth) first frames of our system ranked list output.

Blind taste evaluation questionnaires are based on Grauman et al. [35], which our team had previously [1]. They ask reviewers to answer simple questions comparing different solutions, without them knowing how each result was generated.

We made some improvements to the questionnaries previously used in [1] in order to add more randomness on presentation, to avoid influence of other summaries, and to automatically create them. A brand new online platform was developed to collect these data in an interactive way. Questionnaires are auto-generated from the resulting folders obtained by our system. Each pair of questionnaires is constructed by each dataset folder, which contains all four experiments plus ground-truth, uniform sampling and full event version. To avoid repetitions of summaries a file *show.txt*, which contains the summaries to be shown, is read.

The first round of questionnaires developed (Figure 12) compared the four configurations. For each event, the questionnaire works as follows:

1) First, all the frames of an event are shown to the evaluators.
2) Secondly, the evaluator can watch each summaries and is asked, *Is the summary representative of the event?*. In each event, the order of presented summaries changes randomly. To avoid cases of user answering the question influenced by other summaries, he/she needs to hover mouse above the summary. As our four configurations often give exactly the same summary, only one copy of the summary is presented.
3) Finally, when all four questions are answered, all summaries become visible and the user is asked *Which summary he/she prefers*.

The second round of questionnaires (screen-shot available on figure 13) asked the users to grade each summary comparing to others in order to establish how good the best results of the first round were. The summaries presented were: Best configuration of round one, ground-truth annotated by them months before and a uniform sampling summary. A uniform sampling summary takes frames equi-distributed on time. For example, if we have a 10 frames event and we want to summarize in 5 frames, we will consequently take one of every two frames.

For each event, the questionnaire works as follows:

1) First, all the frames of an event are shown to the evaluators.
2) Secondly, the user watches all summaries and is asked to *grade the visual summary from 1 (worse) to 5 (best).* In each event, the order of presented summaries changes randomly.
3) Finally, as a petition of doctors a ”comments” field is added.

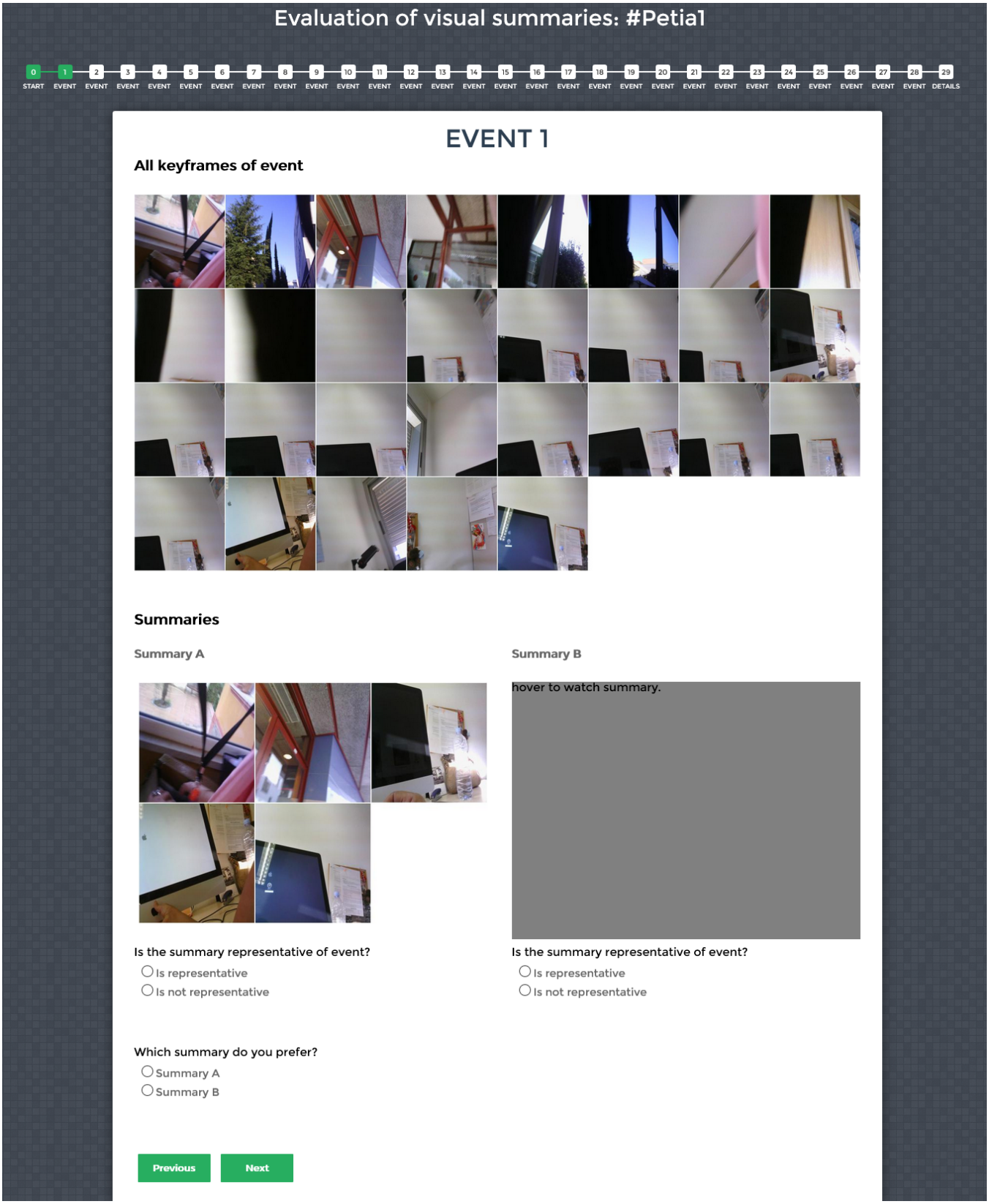

Fig. 12: Screen shot of a first round of online questionnaire presented to doctors corresponding to the first event of the ”Petia1” dataset. Only two summaries are presented, because three of the four configurations give the same result.

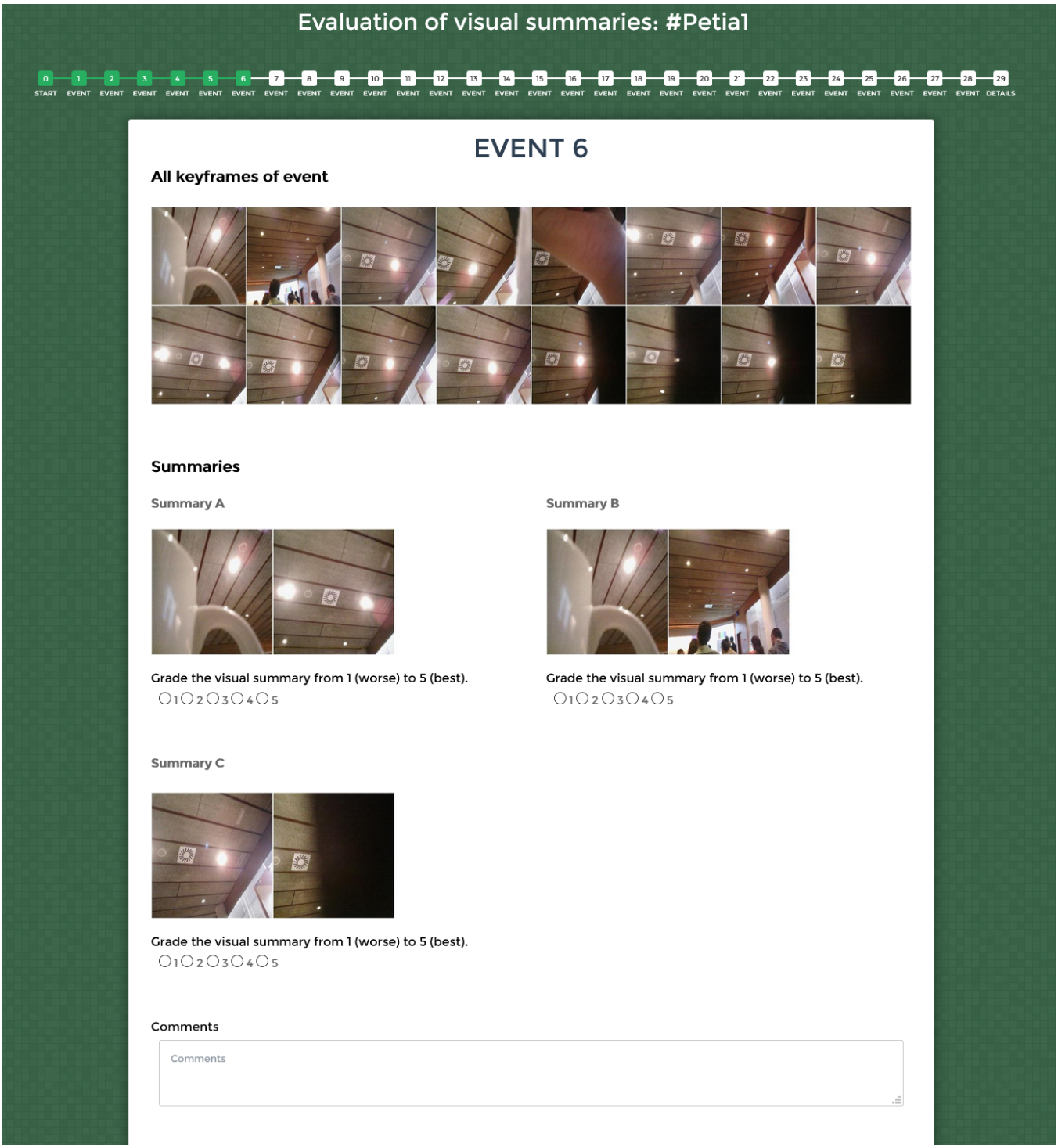

Fig. 13: Screen shot of a second round of online questionnaire presented to doctors corresponding to sixth event of the ”Petia1” dataset.

V. RESULTS

In the experiments section (IV), we discussed about both types of evaluation systems, we have: on one hand doctors' manual evaluation, and on the other hand, an intermediate soft evaluation using MNSMS.

A. Intermediate results

For the **prefiltering** stage, three configurations were assessed (Figure 14). All proposed approaches increase the performance, so we can conclude that this stage is working in the correct way. If we focus on differences of each method, deep learning informativeness strategy on prefiltering sightly improves the basic thresholding results.

In each event, the range of blur and gray mean values (which will allow us to filter dark and burned images) is different. Assigning a threshold for all database is hard due to the range changes. These range changes are mostly due to illumination of each event scene. A better approach for basic thresholding could be to make the threshold relative to the event values or to consider different light models: *outdoor day*, *outdoor night*, *indoor*. Models would need to be detected and a different threshold configuration will be applied.

Regarding the deep learning informativeness strategy, we need to be aware that only frames, where informativeness network is absolutely confident, are removed. Learning somehow a more restrictive threshold for this network would increase results. It is necessary to not forget that in the prefiltering stage images are removed, so we need to keep a high recall.

For the **saliency relevance method**, two configurations were assessed. Both experiments gave very similar results with an AUC around 0.78. Experimental results have shown us that centering a Gaussian did not improve performance. Reviewing Saliency maps literature, we realized that learned ground-truth also has been filtered by a Gaussian distribution, so the SalNet network we are using, has learned to filter the image in this way. Adding an extra Gaussian does not modify results, because relevant information in saliency maps is already at the center of the image.

For the **object relevance method**, two options were tested. As we can see in Figure 15, the LSDA object detector method clearly outperformed the MCG object candidates method. Our intuition of why the object detector clearly beats the object candidates is because objects detected by LSDA are few clearly objects. MCG gives thousands of possible objects, when summing all scores, we introduce noise produced by unlikely objects. The strength of LSDA is the NMS step, where only clear objects are kept.

For the **affective relevance method**, four solutions were considered. We can see in Figure 18 that the sentiment that gives better results is negative, followed by positive and finally extreme sentiments. Best configuration is worse than a random relevance. For this reason, we discarded this affective relevance for the final configuration. We hypothesize that this bad result is due to the kind of images used in the trained CNN, which came from twitter images. Egocentric images are very different from those images. Some domain adaptation or a specific dataset for sentiment in egocentric vision would be needed to increase performance.

For the **relevance fusion**, two experiments have been proposed. We can see in Figure 16 that using a normalization by rank clearly outperforms normalization by score. Our suspicions are that different methods make different distribution of scores, so despite normalization scores are grouped, different methods are not fairly fused. Rank normalization instead forces methods to weight equally.

For the **LSDA used as similarity**, five experiments have been proposed. We can see in Figure 17 that using MNSMS, we can obtain the same conclusions as using Mean Average Precision measure. *LSDA mean similarity* gave the worst result as expected. Also, as expected, *LSDA positive max* outperforms *LSDA max*. Although we were expecting that the original LSDA gives the worse results, it has has held well, having a better performance than *LSDA max* and *LSDA mean*. Surprisingly, *LSDA with only spatial NMS* and *LSDA positive max* gave exactly the same results. *LSDA spatial NMS* is doing the same operation as maximum, the difference between both algorithms is that in *LSDA with only spatial NMS* repeated objects sum scores and the *LSDA positive max* repeated objects takes the score of the maximum of them. In the database, repeated objects inside the same frame are uncommon. In the few situations, it happens, we suppose that it does not do any essential difference. Hence, for the final configuration, we are going to use the *LSDA with only spatial NMS*.

B. Final results

The psychologist team was asked to answer two rounds of questionnaires. In the first round, they were asked to compare the four diversity proposed final configurations (diversity based on ImageNet, Places, LSDA or the fusion of all three similarity measures). In the second round, they were asked to compare the best configuration to their own ground-truth and to a temporal uniform sampling summaries.

The first round of questionnaires had two questions for each event, first doctors decided if each summary presented was representative and then they chosed the preferred one.

The following results of representativity were obtained with the answers of the question *Is the summary representative of the event?*:

ImageNet	Places	LSDA	All fused
95,74%	94,33%	93,62%	94,33%

TABLE II: Percentage of summaries chosen as representative.

All four systems results are considered as representative. The punctuation for each one is above 90%. Therefore, we can consider that our four configurations are correctly summarizing events. The diversity based on ImageNet configuration has the higher performance, narrowly overcoming other configurations.

The second question asked was: *Which summary do you prefer?* and evaluators were able to choose their preferred questionnaire. Just one answer was allowed, but configurations, which gave the same results were grouped. Therefore, sometimes the selected answer voted more than one summary. Summing all four results, we can realize that more than a hundred per cent is obtained. This happens because equal summaries are only presented once, when it happens although user can just vote one summary all equal summaries increment score (all get voted).

ImageNet	Places	LSDA	All fused
59,57%	53,19%	55,32%	54,61%

TABLE III: Percentage of summaries chosen as preferred.

As in the first question, diversity based on ImageNet overcomes the rest of configurations followed by LSDA, Fusion and Places. Curiously, LSDA, which gave the worst result in the first question, achieved second position. The short margin between results shows that all similarity metrics are working similar. Fusing similarities has not increased the performance, results obtained by fusion stays in an intermediate point of the other distances. This reflects that the approach followed for fusing distances, is not bringing improvements or similarity measures used are too similar.

Second round of evaluation aimed to determine how far to the ground-truth and to the uniform sampling our results were. Annotators were asked to *grade each visual summary from 1 (worse) to 5 (best)*. Results obtained for each data-set are:

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**

TABLE IV: Mean Opinion Score for ImageNet, ground-truth and uniform sampling summaries.

Results obtained by uniform sampling (3,99/5) are truly commendable, since they are qualified as *good*. The results obtained with ImageNet are promising, because are closer to the ground-truth than to the uniform sampling.

Doctors have shown coherence with their ground-truth giving a 4,94 of Mean Opinion Score. Results are consistent in all datasets except in the "MAngeles1" datase, where surprisingly our system gave a better performance. Our suspicions are that in "MAngeles1" there are some very dark events from a Easter processions where the number of ground-truth keyframes are very large and it is very difficult to choose the better summary, as our algorithm has prefiltered the darkest ones the overall impression is better although content does not.

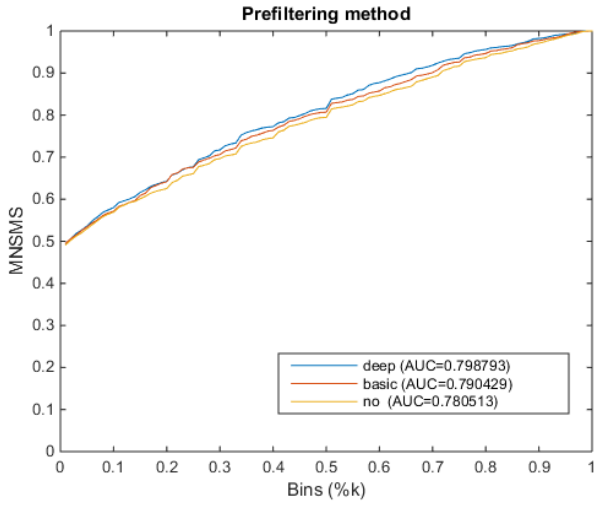

Fig. 14: MNSMS graph of prefiltering methods.

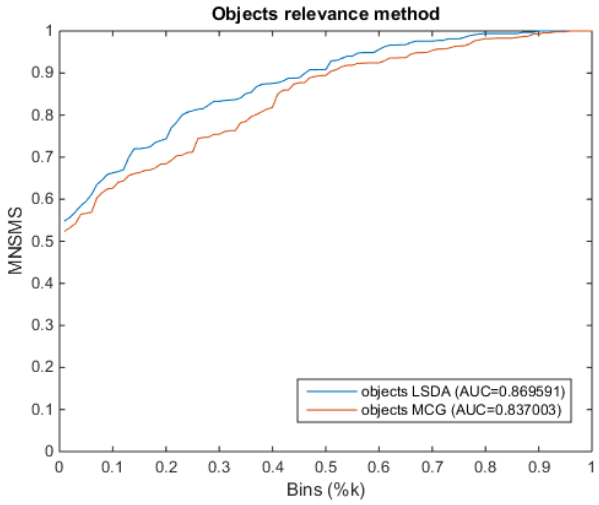

Fig. 15: MNSMS graph of object relevance methods.

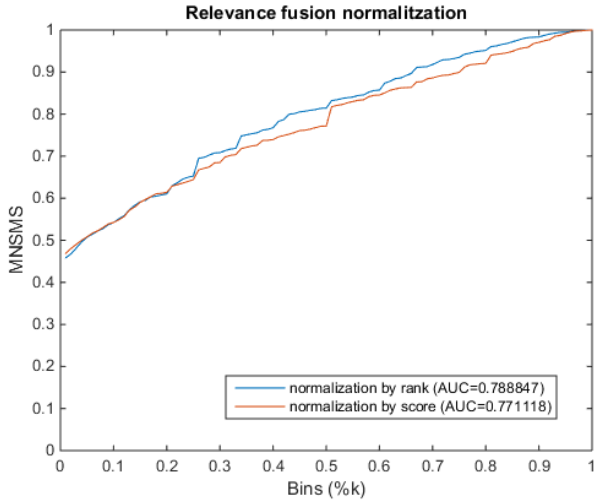

Fig. 16: MNSMS graph of relevance fusion normalization approaches.

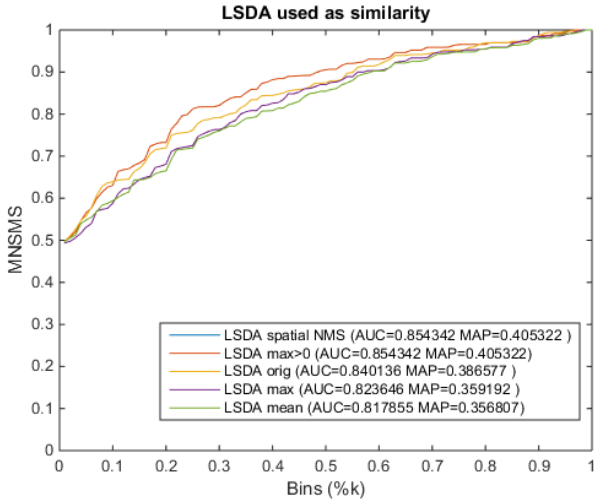

Fig. 17: MNSMS graph of LSDA used as similarity in distance different configurations.

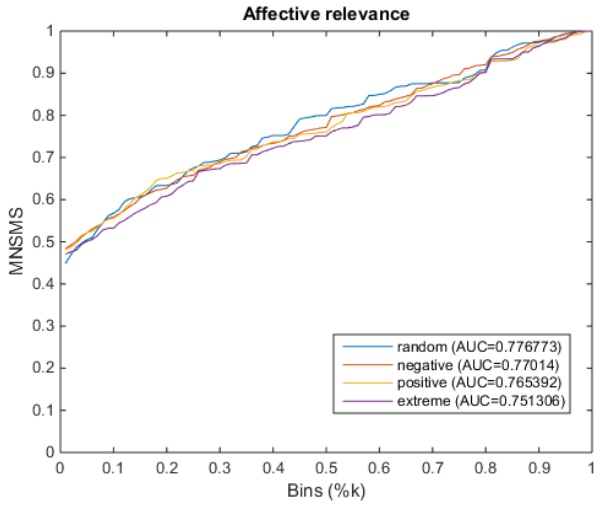

Fig. 18: MNSMS graph of affective relevance configurations.

C. Mediaeval 2015 benchmark participation

Part of the system of this thesis was also tested in the 2015 Retrieving Diverse Social Images Task from the scientific MediaEval benchmark. The participation has involved also Markus Seidl and Matthias Zeppelzauer from St. Pölten University of Applied Sciences who provides the textual analysis of this work. I presented our results and paper [2] at the workshop in Wurzen (Germany) on Monday 14 September 2015 thanks to a student travel grant sponsored by MediaEval.

1) Goal: The MediaEval considers a tourist use case scenario, where a person tries to find more information about a place, she is potentially visiting. To decide whether this place is worth visiting, the person is interested in getting a complete visual summary of that location. The goal of the challenge is to provide a ranked list of Flickr photos for a predefined set of queries. The refined list should be both relevant to the query and also diverse.

For the task, we were expected to submit 5 runs:

- *Run1* had to be computed by visual information only.
- *Run2* - text information only.
- *Run3* - text and visual fused.
- *Run4* - user annotation credibility descriptors (which we have not exploded).
- And in *Run5* - everything was allowed.

2) Method: We applied a methodology of four steps in all our submitted runs very similar to the system described in the thesis:

i) Ranking by relevance: A relevance score for each image is estimated by either using visual or textual information.
ii) Filtering of irrelevant images: Only a percentage of the top ranked images by relevance are considered in later steps. In the multimodal runs, the relevance scores for the visual and textual modalities are linearly normalized and fused by averaging.
iii) Feature and distance computation: Visual and/or textual features are extracted for each image, and the similarity between each pair computed.
iv) Reranking by diversity: An iterative algorithm selects the most different image with respect to all previously selected ones. The similarity is always assessed by averaging the considered visual and textual features. Iterations start by adding the most relevant image as the first element of the re-ranked list.

The system differed from the egocentric system basically in the filtering step. The ranking on relevance makes the decision of discarding images (considering the top frames as relevant), the thesis system instead uses a different criterion to exclude images.

The visual information is analyzed with Convolutional Neural Networks (CNN) [31] with two different purposes:

i) *Ranking by relevance:* Relevance criteria used differ from the ones used on the egocentric method due that the application changes. The network used instead is the Informativeness Network, already presented in the prefiltering stage of the egocentric system.
ii) *Feature and distance computation:* The Places network and ImageNet networks presented in our similarity measures section (III-D) were used to diversify. In the ImageNet network, we use fully connected seventh (*fc7*) layer as in the egocentric case. In the Places network instead, we use *fc8* layer.

The textual information for relevance and distance was obtained by the people from St. Pölten University of Applied Sciences. In the annexes (VI-B) of this thesis, you can find how they computed it.

3) Experiments: The experimental setup was mostly defined by the 2015 MediaEval Retrieving Diverse Images Task, which provided a dataset partitioned into development (devset) and test (testset), two types of queries (single and multi-topic), and standardized and complementary evaluation metrics: Precision at 20 (*P@20*), Cluster Recall at 20 (*CR@20*) and F1-score at 20 (*F1@20*).

The portion of images to be filtered in Step 2 was learned by measuring the evolution of the final F1-score for different percentages. From *Runs1* to *3*, the best results were obtained by keeping the top 20% of images, while for *Run5*, the best value was 15%.

Modality	Visual	Text	Multi	Multi
devset	**Run 1**	**Run 2**	**Run 3**	**Run 5**
P@20	0.756	0.802	0.836	**0.847**
CR@20	0.416	0.419	**0.452**	0.447
F1@20	0.530	0.543	**0.578**	0.577
testset (single)	**Run 1**	**Run 2**	**Run 3**	**Run 5**
P@20	0.705	0.6819	**0.749**	0.733
CR@20	0.423	0.383	**0.431**	0.412
F1@20	0.519	0.478	**0.533**	0.513
testset (multi)	**Run 1**	**Run 2**	**Run 3**	**Run 5**
P@20	0.593	**0.724**	0.627	0.621
CR@20	0.403	0.372	**0.414**	0.397
F1@20	0.463	0.47	**0.482**	0.464
testset (overall)	**Run 1**	**Run 2**	**Run 3**	**Run 5**
P@20	0.649	**0.703**	0.688	0.677
CR@20	0.413	0.378	**0.422**	0.405
F1@20	0.491	0.474	**0.508**	0.489

TABLE V: Precision, Recall and F1-Scores obtained on each run with $N = 20$ on the devset, and the testset (single-topic, multi-topic and overall).

4) Results: Table V presents the results obtained in four different configurations: using visual information only (*Run 1*), using textual data only (*Run2*), and using the best combination of textual and visual data (*Run3*). An additional *Run5* considers multimodal information only for relevance filtering (Step 2) and purely visual information for diversity reranking (Step 4). We skipped *Run4*, because in preliminary experiments the credibility data was not helpful. Rows 2 to 5 present results on the devset for single-topic queries, while rows 6 to row 13 include the results on the testset for the single-topic and multi-topic queries.

From results on the dev and test sets, we draw several conclusions:

Multi-topic queries seem to be more difficult to diversify than single-topic queries. A reason may be that multi-topic queries are more general and contain more heterogeneous content. Considering the fact that our method was trained on single-topic queries only, the results for the multi-topic queries are, however, still promising.

There is no clear winner between textual and visual information (*Runs1* and *2*). The multimodal combination, however, clearly improves performance (*Runs3* and *5*). Additionally, results indicate that using multimodal processing at all stages (*Run 3*) is better than using multimodal processing only during the relevance ranking (*Run 5*).

The results of all the teams can be seen in table VI:

Team	**Run 1**	**Run 2**	**Run 3**	**Run 4**	**Run 5**
USEMP	0.542	0.549	0.544	-	0.542
UPC-UB-STP	0.491	0.474	0.508	-	0.489
IITK-ETH	0.545	0.521	0.539	-	-
LAPI	0.516	0.523	0.506	0.533	0.499
DISI-DIEE	0.496	0.475	0.482	0.503	0.548
UNED-UV	0.488	0.548	0.505	0.499	0.538
USC	0.504	0.517	0.509	-	-
OHSU	0.46	0.42	-	0.46	0.41
Recod	0.541	0.569	0.559	0.553	**0.585**
PKU	0.496	0.503	0.510	0.410	0.524
MIS	0.543	0.511	0.547	-	0.509
DClab	0.478	0.457	0.460	0.456	-
imcube	0.496	0.520	0.513	-	0.435
TUW	0.545	0.497	0.573	0.560	0.549

TABLE VI: F1-score at 20 for all runs for each team in overall testset.

Our results are comparable to other groups. Only by adding credibility information (*Run4*) groups get an increase of performance compared to us. So, this is a direction, we may should go in future.

By combining all text and all visual descriptors (*Run3*), mostly all groups get a slightly decrease of performance. Our team instead increases performance. This shows that our fusion of features is quite efficient.

VI. CONCLUSIONS

A. Conclusions

The main goal for this work was to summarize a set of images using jointly relevance and diversity criteria. As we had seen in the Results section, this objective has been completed with satisfactory results outperforming uniform temporal sampling solutions. Medical physicians validated the proposed solution with a Mean Opinion Score of 4.6 out of 5.00.

Two contributions have been presented in this thesis: A new soft-metric to assess subjective results and a semantic diversity. The soft-metric (Mean Normalized Sum of Max Similarities) allowed us to do intermediate evaluation, which allows to adjust variables of configurations without needing the team of psychologists. Although the semantic diversity (based on LSDA) did not beat ImageNet-based performance, results stayed very close to it.

Also a an automatic web page for online evaluation of summaries has been implemented for this project. This interface takes care of the influence when presenting summaries. It will help GPI-UPC and BCN-PCL to present future results to IR3C-CST psychologists or to other evaluators.

Our work has been tested in two applications: In memory reinforcement for mild-dementia patients, which gave the motivation of this project and re-ranking in image retrieval by participating in the 2015 Retrieving Diverse Social Images Task from the scientific MediaEval benchmark.

There is still some more work to go, to construct the final usable application available in Cognitive Stimulation Theraphy [10], automatically summarizing mild-dementia lifelogs. Yet, for now on, the system can be used to speed-up manual selection of images, working in a semi-supervised way.

The results of this work will be included in a future submission to the Special Issue on: Wearable and Ego-vision Systems for Augmented Experience of the journal IEEE Transactions on Human-Machine Systems. All the code implemented for this work will be publicly available in the GPI website. [2].

B. Future work

In order to improve the results obtained in this work, the following possible improvements arise.

In the first place, it is necessary to further elaborate the relevance criterion. In this project, we focused on two criteria: answering the questions *What?* and *Who?*. In frame level, relevance concept for doctors contained five additional criteria. As future work, we plan to deepen on the rest of criteria.

Doctors also commented higher levels of abstraction, when talking about relevance. The only work done in this thesis in higher levels has been in diversity. Studying relations concepts from different keyframes can open another research line. Also, a lot of work can be done, when answering the questions *What?* and *Who?*. We can try to directly recognize activities and interactions as Ghosh et al [39].

Using different models for different illumination can help us to increase precision of algorithms used. We propose to divide events in *indoor*, *outdoor day* and *outdoor night*. Also, different relevance models can be applied. Observing our final results, we saw decreasing diversity performance on moving events (typically, outdoor) events, frame dissimilarity gets higher when moving, approaches as random walk can be used in this kind of scenes.

Fusion in both relevance and diversity methods can be enhanced. Our intuition is that in relevance fusion, in order to take full advantage of each cue, a determination of which kind of information is binged, would be needed, and a deeper learning of scores weighting is also needed.

In the final results section (V-B), we have seen that diversity fusion did not increased performance. We propose several modifications to struggle the diversity fusion. An idea that can be considered as future work is to subtract the maximum similarity instead to the pondered sum. This will ensure maximum diversity for all similarities. Re-ranking by Soft Max Diversity Fusion, the third step equation (equation 8) would be as follows:

$$\overline{R}_s(k) = R_s(k) - \max_{s=1}^{S} \max_{n=1}^{N} sim_s(R(k), D(n)) \tag{12}$$

Another important research line is to determine automatically the summary length. Our intuition is that we can obtain this information from the diversity measures. Graphs in figures 19 and 20 show the evolution of the similarity in two events of the "Petia1" and "Petia2" datasets. This graph has been computed using equation 13.

$$S(k) = \sum_{n=1}^{k} \min_{n>k}(R(k), R(n)) \tag{13}$$

[2]https://imatge.upc.edu/web/publications/semantic-and-diverse-summarization-egocentric-photo-events-software

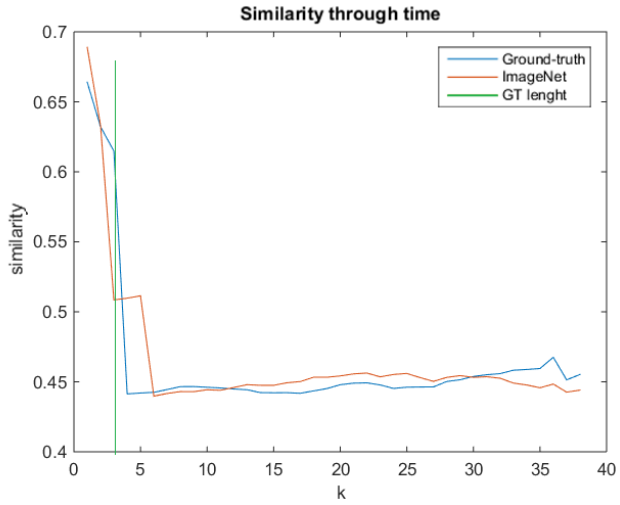

Fig. 19: Similarity through time of event 2 on Petia2 dataset.

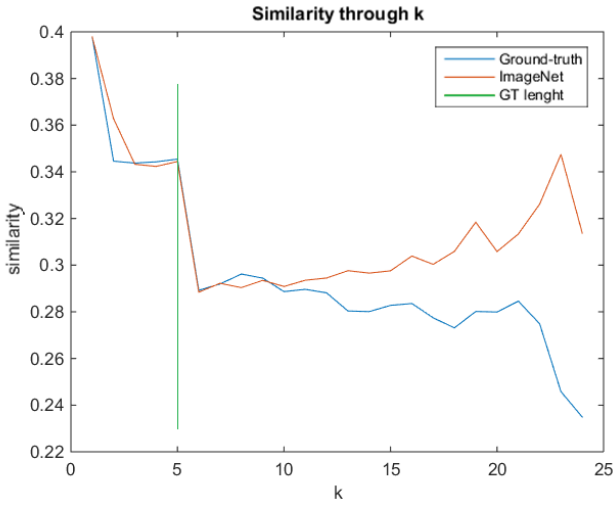

Fig. 20: Similarity through time of event 1 on Petia 1 dataset.

In figures 19 and 20, we can see that the fall down of similarity is given few frames after the ground-truth ideal length. If we take into account that some dissimilar frames do not belong to the summary due to the fact that they are unwanted images, we assume that it may be possible to estimate the length of a summary.

APPENDIX A
MEDIAEVAL 2015 BENCHMARK WORKING NOTES PAPER

UPC-UB-STP @ MediaEval 2015 Diversity Task: Iterative Reranking of Relevant Images

Aniol Lidon
Xavier Giró-i-Nieto
Universitat Politècnica de Catalunya
Barcelona, Catalonia/Spain
xavier.giro@upc.edu

Marc Bolaños
Petia Radeva
Universitat de Barcelona
Barcelona, Catalonia/Spain
marc.bolanos@ub.edu

Markus Seidl
Matthias Zeppelzauer
St. Pölten University of Applied Sciences
St. Pölten, Austria
m.zeppelzauer@fhstp.ac.at

ABSTRACT

This paper presents the results of the UPC-UB-STP team in the 2015 MediaEval Retrieving Diverse Images Task. The goal of the challenge is to provide a ranked list of Flickr photos for a predefined set of queries. Our approach firstly generates a ranking of images based on a query-independent estimation of its relevance. Only top results are kept and iteratively re-ranked based on their intra-similarity to introduce diversity.

1. INTRODUCTION

The diversification of search results is an important factor to improve the usability of visual retrieval engines. This motivates the 2015 MediaEval Retrieving Diverse Images Task [8], which defines the scientific benchmark targeted in this paper. The proposed methodology solves the trade-off between relevance and diversity by firstly filtering results based on a learned relevance classifier, and secondly building a diverse reranked list following an iterative scheme.

The first challenge in our system is filtering irrelevant images, as suggested in [2]. Relevance is a very abstract concept with a high subjectivity involved. Similar problems have been addressed in the visual domain, as for memorability [10] or interestingness [16]. In both cases, a crowdsourced task was organised to collect a large amount of human annotations used to train a classifier based on visual features.

The second challenge to address is the diversity in the ranked list. A seminar work from 1998 [1] introduced diversity in addition to relevance for text retrieval, a concept that was later ported to image [17, 4, 19] and video retrieval [7, 6]. Different features have been used for this purpose, both textual (e.g. tags [20]), visual (e.g. convolutional neural networks [18]), or multimodal fusion [5].

2. METHODOLOGY

A generic and easily extensible methodology of four steps has been applied in all our submitted runs. While steps 2 and 4 apply to all runs, steps 1 and 3 contain particularities for visual and textual processing.

1) Ranking by relevance: A relevance score for each image is estimated by either using visual or textual information (see details in Section 2.1 and 2.2 respectively).

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

2) Filtering of irrelevant images: Only a percentage of the top ranked images by relevance are considered in later steps. In the multimodal runs, the relevance scores for the visual and textual modalities are linearly normalized and fused by averaging.

3) Feature and distance computation: Visual and/or textual features are extracted for each image, and the similarity between each pair computed.

4) Reranking by diversity: An iterative algorithm selects the most different image with respect to all previously selected ones. The similarity is always assessed by averaging the considered visual and textual features. Iterations start by adding the most relevant image as the first element of the reranked list.

2.1 Visual data

The visual information was analyzed with Convolutional Neural Networks (CNN) [13, 12] with two different approaches:

1) Ranking by relevance: A Relevance CNN was created based on HybridNet [22], a CNN trained with objects from the ImageNet dataset [3] and locations from the Places dataset [22]. HybridNet was fine-tuned in two classes: *relevant* and *irrelevant*, as labeled by human annotators.

3) Feature and distance computation: The fully connected layers *fc7* from a CNN trained on ImageNet [11], and the fully connected layer *fc8* from HybridNet [22] were used as feature vectors [14].

2.2 Textual data

1) Ranking by relevance: For each query, we generate a textual term model in an unsupervised manner from all images returned for this query. We first remove stopwords, words with numeric and special characters and words of length ≤ 4. Next, we select the most representative terms by retaining only those terms where the term frequency (TF_q) is higher than the document frequency (DF_q) for the query q. For each term in the model we store the TF_q as a weight. Once this model has been established, we map the textual descriptions of the images to the model of the query. For each image only terms that appear also in the query model are retained. For each remaining term we retrieve the TF_i for the corresponding ith image and build a feature vector. To compute a relevance score s_i for an image, we compute the cosine similarity sim_i between the query model and a given image feature vector. Additionally, we add the inverse original Flickr rank r_i of the image to the score, yielding a final textual relevance score of $s_i = sim_i + (1/r_i)$ for im-

age i. This computation is inspired by that of [21] with the difference that we use TF instead of TFIDF in the scoring function which showed to be more expressive in our experiments.

3) Feature and distance computation: Diversity re-ranking requires the similarity comparison of all relevant images for a query. For a given image, we first align its terms to the query model. Next, we compute their TFIDF weights (TF_i/DF_i) [15, 23]. Terms from the query model that do not occur in the image's descriptions get a weight of zero. The resulting feature vectors are compared with the cosine metric in diversity re-ranking.

3. EXPERIMENTAL SETUP

The experimental setup is mostly defined by the 2015 MediaEval Retrieving Diverse Images Task, which provides a dataset partitioned into development (devset) and test (testset), two types of queries (single and multi-topic), and standardized and complementary evaluation metrics: Precision at 20 (*P@20*), Cluster Recall at 20 (*CR@20*) and F1-score at 20 (*F1@20*). The reader is referred to the task overview paper [8] to learn the details of the problem.

The Relevance CNN described in Section 2.1 was trained with a 2-fold cross validation, each split containing one half of the devset queries. For both splits we stopped after 2,000 iterations, when the validation accuracy was the highest one (76% and 75% respectively). When applying the best methods' parameters on the testset, we used all the dev data and fine-tuned the network stopping after 4,500 iterations, when the training loss was minimum.

The portion of images to be filtered in Step 2 was learned by measuring the evolution of the final F1-score for different percentages. From *Runs 1* to *3* the best results where obtained by keeping the top 20% of images, while for *Run 5* the best value was 15%.

4. RESULTS

Table 1 presents the results obtained in four different configurations: using visual information only (*Run 1*), using textual data only (*Run 2*), and using the best combination of textual and visual data (*Run 3*). An additional *Run 5* considers multimodal information only for relevance filtering (Step 2) and purely visual information for diversity reranking (Step 4). Rows 2 to 5 presents results on the devset for single-topic queries, while rows 6 to row 13 include the results on the testset for the single-topic and multi-topic queries. The overall results can be found in Rows 14 to 17.

Figure 1 plots the Precision, Cluster Recall and F1-Score curves depending on the amount of N top ranked images considered in the evaluation, averaged over all queries on our best run (Run 3).

Modality	Visual	Text	Multi	Multi
devset	**Run 1**	**Run 2**	**Run 3**	**Run 5**
P@20	0.756	0.802	0.836	**0.847**
CR@20	0.416	0.419	**0.452**	0.447
F1@20	0.530	0.543	**0.578**	0.577
testset (single)	**Run 1**	**Run 2**	**Run 3**	**Run 5**
P@20	0.705	0.6819	**0.749**	0.733
CR@20	0.423	0.383	**0.431**	0.412
F1@20	0.519	0.478	**0.533**	0.513
testset (multi)	**Run 1**	**Run 2**	**Run 3**	**Run 5**
P@20	0.593	**0.724**	0.627	0.621
CR@20	0.403	0.372	**0.414**	0.397
F1@20	0.463	0.47	**0.482**	0.464
testset (overall)	**Run 1**	**Run 2**	**Run 3**	**Run 5**
P@20	0.649	**0.703**	0.688	0.677
CR@20	0.413	0.378	**0.422**	0.405
F1@20	0.491	0.474	**0.508**	0.489

Table 1: Precision, Recall and F1-Scores obtained on each run with $N = 20$ on the devset, and the testset (single-topic, multi-topic and overall).

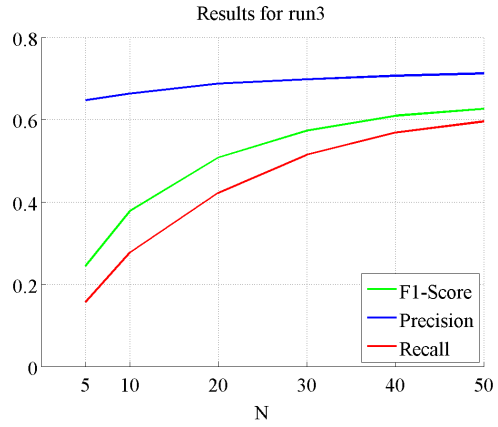

Figure 1: Overall Precision, Recall and F1-score curves for different cutoffs N of top ranked images on all testset queries.

5. CONCLUSIONS

The trade-off between relevance and diversity has been targeted in this work with relevance-based filtering and a posterior iterative process to introduce diversity. The final results, presented in Table 1, are comparable to the state of the art on the devset [9], and achieve up to a F1@20 of 0.508 on the testset.

Multi-topic queries seem to be more difficult to diversify than single-topic queries. A reason may be that multi-topic queries are more general and contain more heterogeneous content. Considering the fact that our method was trained on single-topic queries only, the results for the multi-topic queries are, however, still promising.

It is remarkable that increasing the number of N of retrieved images increases both, recall and precision (and not only recall as one would expect in a typical retrieval scenario), as shown in Figure 1. This indicates that the relevance ranking obtained by our method is accurate (at least for $N \leq 50$).

There is no clear winner between textual and visual information (*Runs 1* and *2*). The multimodal combination, however, clearly improves performance (*Runs 3* and *5*). Additionally, results indicate that using multimodal processing at all stages (*Run 3*) is better than using multimodal processing only during the relevance ranking (*Run 5*).

6. REFERENCES

[1] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM, 1998.

[2] D.-T. Dang-Nguyen, L. Piras, G. Giacinto, G. Boato, and F. G. De Natale. A hybrid approach for retrieving diverse social images of landmarks. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, pages 1–6. IEEE, 2015.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[4] T. Deselaers, T. Gass, P. Dreuw, and H. Ney. Jointly optimising relevance and diversity in image retrieval. In *Proceedings of the ACM international conference on image and video retrieval*, page 39. ACM, 2009.

[5] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, and X. Wu. Visual-textual joint relevance learning for tag-based social image search. *Image Processing, IEEE Transactions on*, 22(1):363–376, 2013.

[6] X. Giro-i Nieto, M. Alfaro, and F. Marques. Diversity ranking for video retrieval from a broadcaster archive. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 56. ACM, 2011.

[7] M. Halvey, P. Punitha, D. Hannah, R. Villa, F. Hopfgartner, A. Goyal, and J. M. Jose. Diversity, assortment, dissimilarity, variety: A study of diversity measures using low level features for video retrieval. In *Advances in Information Retrieval*, pages 126–137. Springer, 2009.

[8] B. Ionescu, A. L. Gınsca, B. Boteanu, A. Popescu, M. Lupu, and H. Müller. Retrieving diverse social images at mediaeval 2015: Challenge, dataset and evaluation. In *MediaEval 2015 Workshop, Wurzen, Germany*, 2015.

[9] B. Ionescu, A. Popescu, M. Lupu, A. L. Gînsca, B. Boteanu, and H. Müller. Div150cred: A social image retrieval result diversification with user tagging credibility dataset. *ACM Multimedia Systems-MMSys, Portland, Oregon, USA*, 2015.

[10] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(7):1469–1482, 2014.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[14] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 512–519. IEEE, 2014.

[15] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[16] M. Soleymani. The quest for visual interest. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 2015.

[17] K. Song, Y. Tian, W. Gao, and T. Huang. Diversifying the image retrieval results. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 707–710. ACM, 2006.

[18] E. Spyromitros-Xioufis, S. Papadopoulos, A. L. Ginsca, A. Popescu, Y. Kompatsiaris, and I. Vlahavas. Improving diversity in image search via supervised relevance scoring. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 323–330. ACM, 2015.

[19] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol. Visual diversification of image search results. In *Proceedings of the 18th international conference on World wide web*, pages 341–350. ACM, 2009.

[20] R. Van Zwol, V. Murdock, L. Garcia Pueyo, and G. Ramirez. Diversifying image search with user generated content. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 67–74. ACM, 2008.

[21] B. Vandersmissen, A. Tomar, F. Godin, W. De Neve, and R. Van de Walle. Ghent university-iminds at mediaeval 2014 diverse images: Adaptive clustering with deep features. In *MediaEval 2014, Workshop*, 2014.

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

[23] J. Zobel and A. Moffat. Exploring the Similarity Space. *ACM SIGIR Forum*, 32(1):18–34, 1998.

ACKNOWLEDGMENT

The author would like to thank in first place the people from the team, which gave support during this project; without them this project would not be possible to perform: Xavi Giró, Petia Radeva and Marc Bolaños.

In second place, we wold like to thank to the GPI group of UPC for the chance of working with a very powerful software and to their technical support to make that everything works.

We also would like to thank the people of IR3C group to give the psychological view of this project and to make the thesis more consistent.

REFERENCES

[1] R. Mestre, M. Bolaños, E. Talavera, P. Radeva, and X. Giró-i Nieto, "Visual summary of egocentric photostreams by representative keyframes," in *Proc. IEEE International Workshop on Wearable and Ego-vision Systems for Augmented Experience (WEsAX)*, 2015.

[2] A. Lidon, M. Bolaños, M. Seidl, X. Giro-i Nieto, P. radeva, and M. Zeppelzauer, "Upc-ub-stp @ mediaeval 2015 diversity task: Iterative reranking of relevant images," in *MediaEval 2015 Workshop, Wurzen, Germany*, 2015.

[3] S. Mann, "'wearcam' (the wearable camera): personal imaging systems for long-term use in wearable tetherless computer-mediated reality and personal photo/videographic memory prosthesis," in *Wearable Computers, 1998. Digest of Papers. Second International Symposium on*, Oct 1998, pp. 124–131.

[4] W. W. Mayol, B. J. Tordoff, and D. W. Murray, "Wearable visual robots," *Personal Ubiquitous Comput.*, vol. 6, no. 1, pp. 37–48, Jan. 2002. [Online]. Available: http://dx.doi.org/10.1007/s007790200004

[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 *Proceedings of the 8th International Conference on Ubiquitous Computing*, ser. UbiComp'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 177–193.

[6] A. J. Sellen, A. Fogg, M. Aitken, S. Hodges, C. Rother, and K. Wood, "Do life-logging technologies support memory for the past?: An experimental study using sensecam," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '07. New York, NY, USA: ACM, 2007, pp. 81–90. [Online]. Available: http://doi.acm.org/10.1145/1240624.1240636

[7] M. L. Lee and A. K. Dey, "Lifelogging memory appliance for people with episodic memory impairment," in *Proceedings of the 10th International Conference on Ubiquitous Computing*, ser. UbiComp '08. New York, NY, USA: ACM, 2008, pp. 44–53. [Online]. Available: http://doi.acm.org/10.1145/1409635.1409643

[8] P. Piasek, K. Irving, and A. Smeaton, "Sensecam intervention based on cognitive stimulation therapy framework for early-stage dementia," in *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2011 5th International Conference on*, May 2011, pp. 522–525.

[9] P. Piasek, A. F. Smeaton *et al.*, "Using lifelogging to help construct the identity of people with dementia," 2014.

[10] A. SPECTOR, L. THORGRIMSEN, B. WOODS, L. ROYAN, S. DAVIES, M. BUTTERWORTH (deceased), and M. ORRELL, "Efficacy of an evidence-based cognitive stimulation therapy programme for people with dementia," *The British Journal of Psychiatry*, vol. 183, no. 3, pp. 248–254, 2003.

[11] D. Tancharoen, T. Yamasaki, and K. Aizawa, "Practical experience recording and indexing of life log video," in *Proceedings of the 2Nd ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, ser. CARPE '05. New York, NY, USA: ACM, 2005, pp. 61–66. [Online]. Available: http://doi.acm.org/10.1145/1099083.1099092

[12] A. R. Doherty, D. Byrne, A. F. Smeaton, G. J. Jones, and M. Hughes, "Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs," in *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*, ser. CIVR '08. New York, NY, USA: ACM, 2008, pp. 259–268. [Online]. Available: http://doi.acm.org/10.1145/1386352.1386389

[13] M. Blighe, A. Doherty, A. F. Smeaton, and N. E. O'Connor, "Keyframe detection in visual lifelogs," in *Proceedings of the 1st International Conference on PErvasive Technologies Related to Assistive Environments*, ser. PETRA '08. New York, NY, USA: ACM, 2008, pp. 55:1–55:2. [Online]. Available: http://doi.acm.org/10.1145/1389586.1389652

[14] A. Jinda-Apiraksa, J. Machajdik, and R. Sablatnig, "A keyframe selection of lifelog image sequences," *Erasmus Mundus M. Sc. in Visions and Robotics thesis, Vienna University of Technology (TU Wien)*, 2012.

[15] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '98. New York, NY, USA: ACM, 1998, pp. 335–336. [Online]. Available: http://doi.acm.org/10.1145/290941.291025

[16] K. Song, Y. Tian, W. Gao, and T. Huang, "Diversifying the image retrieval results," in *Proceedings of the 14th Annual ACM International Conference on Multimedia*, ser. MULTIMEDIA '06. New York, NY, USA: ACM, 2006, pp. 707–710. [Online]. Available: http://doi.acm.org/10.1145/1180639.1180789

[17] O. Chum, J. Philbin, M. Isard, and A. Zisserman, "Scalable near identical image and shot detection," in *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, ser. CIVR '07. New York, NY, USA: ACM, 2007, pp. 549–556. [Online]. Available: http://doi.acm.org/10.1145/1282280.1282359

[18] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol, "Visual diversification of image search results," in *Proceedings of the 18th International Conference on World Wide Web*, ser. WWW '09. New York, NY, USA: ACM, 2009, pp. 341–350. [Online]. Available: http://doi.acm.org/10.1145/1526709.1526756

[19] T. Deselaers, T. Gass, P. Dreuw, and H. Ney, "Jointly optimising relevance and diversity in image retrieval," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, ser. CIVR '09. New York, NY, USA: ACM, 2009, pp. 39:1–39:8. [Online]. Available: http://doi.acm.org/10.1145/1646396.1646443

[20] B. Ionescu, M. Menéndez, H. Müller, and A. Popescu, "Retrieving diverse social images at mediaeval 2013: Objectives, dataset and evaluation," 2013.

[21] B. Ionescu, A. Popescu, M. Lupu, A. L. Gınsca, and H. Müller, "Retrieving diverse social images at mediaeval 2014: Challenge, dataset and evaluation," in *MediaEval 2014 Workshop, Barcelona, Spain*, 2014.

[22] E. Spyromitros-Xioufis, S. Papadopoulos, Y. Kompatsiaris, and I. Vlahavas, "Socialsensor: Finding diverse images at mediaeval 2014," 2014.

[23] D.-T. Dang-Nguyen, L. Piras, G. Giacinto, G. Boato, and F. De Natale, "Retrieval of diverse images by pre-filtering and hierarchical clustering," *Working Notes of MediaEval*, 2014.

[24] F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas, "The blur effect: perception and estimation with a new no-reference perceptual blur metric," in *Electronic Imaging 2007*. International Society for Optics and Photonics, 2007, pp. 64 920I–64 920I.

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

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[27] P. Arbelaez, J. Pont-Tuset, J. Barron, F. Marqués, and J. Malik, "Multiscale combinatorial grouping," in *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[28] J. Hoffman, S. Guadarrama, E. Tzeng, J. Donahue, R. B. Girshick, T. Darrell, and K. Saenko, "LSDA: large scale detection through adaptation," *CoRR*, vol. abs/1407.5035, 2014. [Online]. Available: http://arxiv.org/abs/1407.5035

[29] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2879–2886.

[30] V. Campos, "Layer-wise cnn surgery for visual sentiment prediction," Master's thesis, 2015.
[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[32] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.
[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.
[34] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.
[35] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 2714–2721.
[36] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
[37] G. A. Mills-Tettey, A. Stentz, and M. B. Dias, "The dynamic hungarian algorithm for the assignment problem with changing costs," 2007.
[38] T. Deselaers and V. Ferrari, "Visual and semantic similarity in imagenet," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1777–1784.
[39] J. Ghosh, Y. J. Lee, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1346–1353.