D5.3

Multimodal content mining and linking framework – v1

v 1.0 / 2015-02-12

Etienne Gadeski (CEA), Elefterios Spyromitros-Xoufis (CERTH), Adrian Popescu (CEA), Symeon Papadopoulos (CERTH), Herve Le Borgne (CEA), Yiannis Kompatsiaris (CERTH)

This deliverable is a report which describes the first version of the USEMP multimodal content mining and linking modules. The report includes, for each module: its functionality, experimental results that support the validity of the choices made, and implementation details that are useful for its integration in the USEMP framework. Similar to the related D5.1 and D5.2 deliverables, the report highlights the importance of multimodal mining modules for the project use cases and multidisciplinary issues related to their development.
Naturally, the modules developed as part of this deliverable rely, to a large extent, on those developed as part of the other two WP5 deliverables. Focus is put on multimodal fusion for: (a) multimodal concept detection, (b) location detection and (c) relevance reranking using personal user feedback. As this deliverable demonstrates, the developed fusion mechanisms can lead to a) considerable improvements in terms of the accuracy of the automatically mined information, and to b) user-adapted multimedia retrieval, both of which are essential requirements for the USEMP system. Multimedia mining modules are available to Online Social Networks (OSNs) and allow them to infer valuable knowledge from raw user data and it in their business models. For instance, large scale concept detection from user's texts and images can give valuable information for the creation of detailed consumer profiles that are exploitable for targeted advertisement. The use of multimedia mining tools by OSNs can have strong impact on users' privacy and, in USEMP, they are integrated in feedback and awareness tools in order to raise users' awareness and control of their personal data.

Project acronym	USEMP
Full title	User Empowerment for Enhanced Online Presence Management
Grant agreement number	611596
Funding scheme	Specific Targeted Research Project (STREP)
Work program topic	Objective ICT-2013.1.7 Future Internet Research Experimentation
Project start date	2013-10-01
Project Duration	36 months

Workpackage	WP5
Deliverable lead org.	CEA
Deliverable type	Prototype
Authors	Etienne Gadeski (CEA) Elefterios Spyromitros-Xoufis (CERTH) Adrian Popescu (CEA) Symeon Papadopoulos (CERTH) Herve Le Borgne (CEA) Yiannis Kompatsiaris (CERTH)
Reviewers	David Lund (HWC) Ali Mohammad Padyab (LTU)
Version	1.0
Status	Final
Dissemination level	RE: Restricted Group
Due date	2015-01-31
Delivery date	2015-02-12

Version	Changes
0.1	Initial draft by Etienne Gadeski and Adrian Popescu
0.2	Input from CERTH
0.3	Refinements from CERTH
0.4	Refinements from CEA
0.5	Refinements based on reviewers' comments
1.0	Final revisions from CERTH and CEA

Table of Contents

1. Introduction

This deliverable provides a description of the USEMP multimodal[1] annotation, retrieval and location detection modules implemented during the first iteration of the project. The introduction first gives an overview of the role of multimodal mining in USEMP, of the research methodology and of multidisciplinary interactions within the project.

The main objectives of the deliverable are:

a) to clarify the usage of multimodal mining modules in the USEMP framework;
b) to show how textual and visual modalities can be effectively combined in order to improve the overall quality of multimedia mining results;
c) to detail the research approaches adopted, including implementation details;
d) to present an evaluation of multimodal mining modules on relevant datasets;
e) to detail how these modules are interfacing with other modules in the USEMP system.

1.1. Multimodal mining and linking in USEMP

The main objective of multimodal mining and linking is to combine text and visual content mining in order to endow the USEMP framework with the capability to **conduct inferences about OSN users' interests and traits based on the multimodal content** they share and interact with. Naturally, multimedia fusion is only doable if a document contains text and image components and, whenever this condition is not met, text mining (D5.1) or visual content mining (D5.2) should be used instead. Inferences over multimodal documents are most often extracted for individual documents, but are subsequently used in other parts of the project, as follows:

- Direct exploiting of multimodal inferences in the platform implemented in WP7;
- Combination with behavioral cues processed as part of the privacy scoring framework (T6.1) and integration in the USEMP platform.

During the first iteration of the project, we prioritized the combination of textual and visual content mining modules while the combination of these cues with behavioral ones studied as part of WP6 will be studied during the second iteration of this deliverable (D5.6). As we mentioned above, the inferences that can be obtained through multimodal mining depend on the available text and visual content mining modules. After an initial analysis of the maturity and capabilities offered by the single modality approaches described in D5.1 (text mining) and D5.2 (visual content mining), the following modules were exploited for multimodal mining:

Text mining:

- TXT_1 (text similarity) – represents texts in a vector space and then exploits these representations to compute similarities.
- TXT_2 (location detection) – extracts an estimation of the most probable coordinates (place name) for an input text.

[1] Here, multimodal refers to content processing techniques that make use of both textual and visual content at the same time.

Visual mining

- *VIS$_1$* (concept detection) – predicts the most probable concepts[2] for an input image.
- *VIS$_2$* (location detection) – extracts an estimation of the most probable coordinates (place name) for an input image.

Multimodal concept annotation Section 2 was implemented for two cases, i.e. absence or presence of textual annotations associated to the target image, noted 1 and 2 below). In the first case, module *VIS$_1$* is used in two settings: (1.a) proposal of probable concepts from a predefined and closed list of concepts using supervised classification and (2.b) proposal of new tags from an open vocabulary determined by the similarity between the input images and those from an annotated reference database. In the second case, i.e. presence of some initial textual annotations, we also explored two possibilities: (2.a) late fusion – separate processing of text annotation and of low-level image descriptors, followed by a fusion of results and (2.b) – late fusion of initial annotations and of results obtained in (1.a) and (1.b).

Modules *TXT$_1$* and *VIS$_2$* are combined in order to obtain improved **multimodal location predictions** (Section 3). Put simply, confidence scores are computed for each modality and, if the probability for one of the two modalities to be right is very high, it is given priority over the other.

The success of privacy enhancement tools is, to a large extent, conditioned by the proposal of appropriate interaction means between the user and the system. In this context, a module which aims at the **reranking of private content (images) based on personal user feedback** is introduced by taking advantage of feature extraction capabilities developed as part of *TXT$_1$* (i.e. bag of words representation of content) and *VIS$_1$* (i.e. low-level visual features associated to images) in a relevance feedback loop. Given a set of results for a private concept of interest that was automatically computed, the user can select one or more images which are considered relevant and a refined set of results is proposed by combining textual and visual features (Section 4).

1.2. Research methodology and contributions

Research on multimodal mining and linking is successively shaped by the conclusions of upstream research from other disciplines: legal studies (WP3), social science (WP4), user studies and system design (WP4, WP2). The links with these research streams are discussed in more details in Subsection 1.3. Naturally, multimodal mining relies on the modules available for text mining (D5.1) and visual content mining (D5.2). The overall objective is to leverage complementary contributions from individual textual and visual modalities in order to improve the obtained inferences. Assuming that the features for the involved modalities are already available, there are two main types of multimodal fusion: (1) early fusion – the features are combined in a common space before performing any further processing (i.e. machine learning for classification or similarity computation for retrieval) and (2) late fusion – a complete processing is performed for each modality and results are combined only at the end, with the visual concept being linked to textual entities. According to recent studies in the field, late fusion has been found to perform better mainly due to the difficulty of early combination of textual and visual modalities. This difficulty is due to the

[2] The term *concept* is an established term in the multimedia analysis and computer vision research communities and is typically associated with a topic, entity, object or theme depicted in an image

differences between the information conveyed by the two channels: while textual descriptions directly provide semantic information that is understandable by the users, visual descriptions convey low level (i.e. pixel related) information that needs to be further processed in order to be understandable by users. In each case, the most effective methods stemming from D5.1 and D5.2 were selected as the basis for the modules, with preference given to reusing modules wherever possible. To assess the usefulness of the proposed prototypes, evaluation was carried out with suitable publicly available datasets or approaches.

Although the multimedia fusion work done during the first iteration of USEMP development cycles relied, to a large extent, on existing NLP and computer vision approaches, we consider that it results in a number of interesting research contributions, including:

- In the case of images with existing sparse annotations, a simpler and more principled combination of image annotations produced by humans and of automatic annotations. This combination is enabled by the creation of a very large number of efficient visual concept classifiers, whose outputs can be seamlessly combined with manual annotations.
- In the case of unlabeled images, an open-vocabulary annotation procedure which exploits the power of Semfeat, the new concept-level feature representation developed as part of D5.2. This type of annotation has the advantage of not being constrained by a predefined concept vocabulary and produces automatic annotations that are close to those proposed by humans.
- A powerful multimedia location prediction framework that leverages the advantages of text-based probabilistic location models (D5.1) and of new convolutional neural network (CNN) based visual location detection approaches (D5.2). Confidence scores are exploited to decide which modality should be used in priority in order to improve overall location detection.
- A principled improvement of an existing multimedia fusion method that is exploited in order to propose an effective way to leverage personal user feedback for improved privacy-aware retrieval, an important component of the USEMP framework. The main improvements come from a supervised definition of the task to solve and from a more principled combination of text and image features.

1.3. Multidisciplinary issues[3]

Multimodal mining operates a combination of text and visual mining results and is thus mainly dealing with approaches from natural language processing, computer vision and machine learning. However, the presented research was considerably shaped by the rest of the USEMP disciplines, and at the same time provides actionable feedback to them. In the following, we provide a concise account of the inter-play between text mining research and the different disciplines of the project.

D5.3 is informed by work done in WP2, WP3, WP4 and WP9 and it provides valuable input for WP6 and WP7. The legal analysis carried out in WP3, and more particularly in T3.6 which deals with the coordination of legal aspects, clarified practical implications of multimodal mining related and were turned into specific requirements that were implemented:

[3] Multidisciplinary issues are, to a large extent, common to all WP5 deliverables and this section has thus similar content in D5.1, D5.2 and D5.3.

- The USEMP end-users should be clearly informed about their rights and obligations when engaging with the platform.
- Processing of personal data should be subjected to a declaration of USEMP work to national Data Protection Agencies.
- Processing of sensitive information, such as user personally identifiable information, should be considered separately and be subjected to a specific declaration.
- Copyright issues should be carefully considered for training data used during USEMP and, more importantly, for any commercial implementation of its results after the end of the project
- Ensuring that all USEMP components have clear IP rights (in case of reusing existing components).

Work on trade secrets and intellectual property done as part of D3.2 explored the tensions between profile representations on the end-user side, within OSNs and created in USEMP and made clear the complex interplay between these actors, as well as their respective rights and obligations.

The use case analysis in D2.1 and the associated requirements defined in D2.2 served as guidelines for the implementation of technical components. In particular, the following system requirements are central here:

- [SR02] [4]"The system may be able to process the information within one second such that the user can make informed decisions on their past data without long delays. In the event data processing is to take longer, a progress bar should be presented. A maximal extent of 10 seconds will be aimed for." This requirement has strong implications in terms of processing speed for the implemented components.
- [SR04] "The system may be able to make best effort associations between data placed onto OSN(s) and the profile attributes which can be inferred from such data." This requirement is a counterpart of [SR02] that focuses on component performance, which should closely follow state of the art developments.
- [SR11] "The system may be able to get fruitful insights on how relevant a user's profile is for different stakeholders." Through inferences made by technical components, the end-users should be able to have insightful information on how her profile is seen by OSNs and, possibly, by other stakeholders.

In D4.1, a comprehensive list of social requirements was established, which offers a user-side view of the expected behavior of the developed USEMP tools. While all requirements are important, the following ones have particular impact on multimedia mining modules:

- Req. 1 asking for more transparency about privacy problems at an institutional level and notably OSNs in this context.
- Req. 2 demanding a backward link between inferences and raw data which generated them to improve the explainability of the automatic decisions made by the system.
- Req. 10 asking for low impact on browser speed of the USEMP plug-in, a requirement which is tightly linked to [SR02] mentioned above.

[4] The requirement notation is the one used in the deliverables that extracted them.

The extensive market analysis done in D9.3 showed that existing privacy enhancing tools and privacy feedback and awareness tools deal mostly with volunteered and/or observed data. A strong opportunity in USEMP is to provide users with a more complete view of how their data could be handled and exploited by OSNs. Another conclusion of D9.3 is that existing text and image mining tools are not tailored for privacy enhancement and, consequently, an adaptation step is needed in order to better satisfy domain requirements. Downstream, insights gained with D5.3 tools can be used both directly in the USEMP interface (D7.2), and as part of the privacy scoring framework created in D6.1, to complement social network mining inferences. For instance, user locations can be extracted from texts and images and can then be displayed directly by the USEMP interface to inform the user about her degree of exposure on a certain privacy dimension (e.g. location). In a more complex functioning mode, multimodal data representations can be combined with social interaction data (such as likes, comments) to improve the quality of predictions.

2. Multimodal concept detection

Building on the concept detection for visual content developed as part of D5.2, we explore ways to combine them with textual annotations associated to multimodal data. Upstream work done in WP2, WP4 and WP6 provided valuable insights about the task to solve. In particular, it became clear that a very large variety of concepts[5], i.e. entities, objects and themes of interest depicted in images, are illustrated in user content shared on OSNs. Consequently, scalability in terms of recognizable concepts should be a core requirement, along with detection accuracy, which central in order to extract reliable knowledge about OSN users and to ensure user's trust in the system. To cope with these requirements it is important to propose approaches which leverage the use of powerful visual concept detectors and of manual image annotations produced by users.

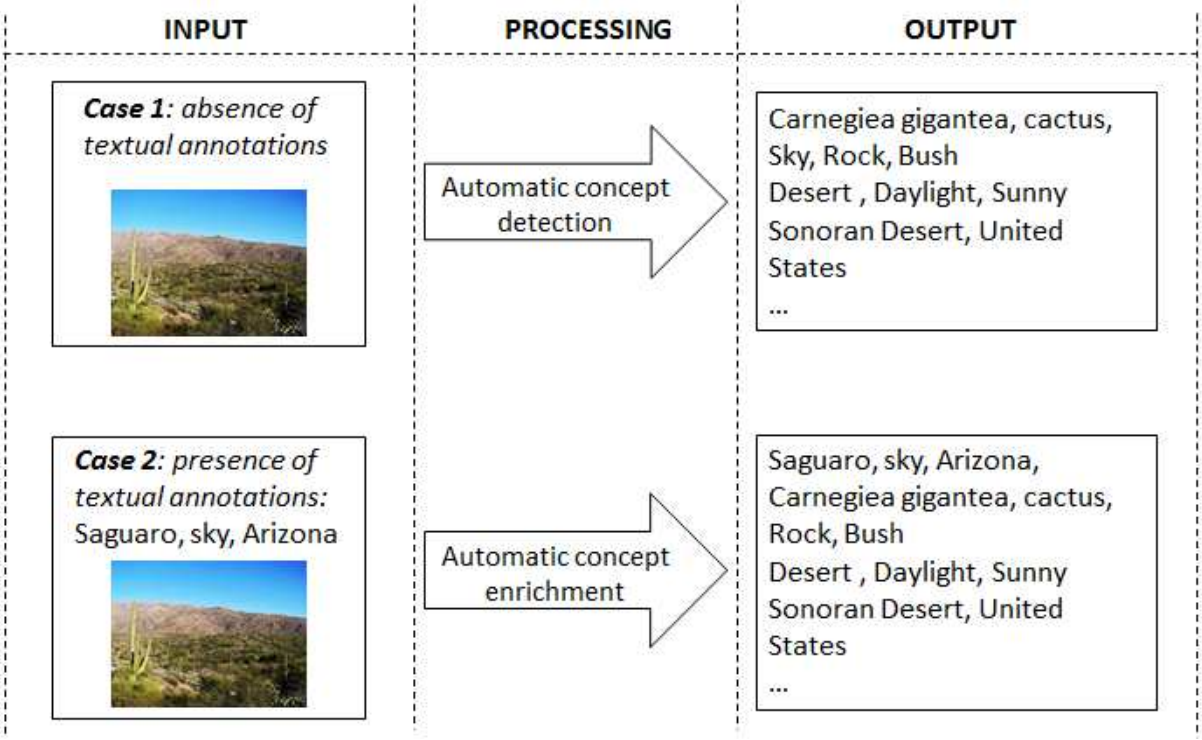

Figure 1. Illustration of the two main types of multimedia concept detection, i.e. in absence or presence of initial manual annotations. ***Case 1*** *predicts output annotations directly from image content and* ***Case 2*** *extends initial annotations provided by users with tags which are associated to its content.*

In Figure 1, we illustrate the two main cases of multimodal concept detection, i.e. absence or presence of initial user annotations that appear for user content processed in USEMP. The first case is naturally more challenging since only the image content is available as input data and, whenever textual annotations are available, they should be exploited. We studied solutions for dealing with both cases presented in Figure 1 and describe them in the following subsections. Similar to the visual concept detection (D5.2), the results of multimodal concept detection can be exploited directly or be combined with other insights, obtained for instance by the social network mining approaches described in D6.1. When exploited directly, the user

[5] The term *concept* is an established term in the multimedia analysis and computer vision research communities and is typically associated with a topic, entity, object or theme depicted in an image.

will receive feedback about privacy-related concepts[6] associated to her profile. For instance, the example image from Figure 1 discloses information about the location of the user and, when aggregated with other location-related images, it could contribute to a thorough location profile.

2.1. Related work

Similar to visual concept detection, multimodal concept detection is usually cast a supervised classification problem in which the low level (i.e. pixel based) and high level (i.e. semantic) descriptions of images are combined in order to take advantage of their complementarities. In absence of textual annotations, the automatic annotations can be created using two main approaches:

- *Detection through learning*, i.e. learning visual representations for a number of concepts and then detecting them in the test images. This method is predominant and is used, for instance, in the ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al., 2014), which operates over a set of 1,000 classes. It is also implemented, for a larger number of concepts by commercial applications such as Clarifai[7]. Its main advantage comes from a robust modeling of each recognizable concept while its main limitation is the fact that only a predefined list of concepts can be recognized. This limitation is important in the USEMP context since users are likely to upload a very large variety of photos. A potential solution to this problem was proposed in D5.2 and described in more details in (Ginsca et al., 2015). It relies on learning a very large number of visual concepts using feature transfer from a CNN learned with 1,000 classes and noisy data from the Web.
- *Detection through similarity*, i.e. predicting tags with a manually labeled reference dataset and a kNN visual similarity calculation (Li et al., 2009). In this approach, image features are extracted offline for the reference dataset and their tags are considered as possible annotations for new images. Given a test image, the most similar images from the reference dataset are computed and their annotations are leveraged to predict annotations for the target image. The main advantage of the method is that it is potentially more expressive than the one based on concept learning due to the fact that any user tag can be used as annotation. The main disadvantage comes from the fact that image-to-image similarities are often less robust than a comparison of a test image with a statistical model abstracted from a pool of concept images.

In presence of manual annotations, the detection task is somewhat simpler since the visual modality can be used to enrich the initial description of the image. For instance, (Znaidia et al., 2012) propose a classifier stacking approach to combine visual and textual modalities using late fusion. Put simply, annotations are predicted first separately for each modality and they are then combined into a final image description.

[6] Here privacy-related concepts should be understood as concepts that are directly tied to the main privacy dimensions determined as part of the privacy scoring framework described in D6.1 and were discussed in more depth in the context of visual content mining (D5.2).
[7] http://www.clarifai.com/

2.2. Method description

2.2.1. Concept detection in absence of textual annotations

An instantiation of multimedia concept detection in absence of textual annotations (based on learning a large number of concepts) was presented and evaluated in D5.2. Here we focus on an instantiation of a similarity based method inspired by (Li et al., 2009). Given the USEMP setting, i.e. very broad conceptual coverage of the photos and large volumes of data to be processed fast, the similarity based annotation method needs to:

- Mine a large-scale and highly diversified reference dataset in order to annotate different types of content.
- Exploit a compact but accurate image representation in order to compute annotations in real-time.

To comply with these requirements, we build on the Semfeat representation of images, introduced in D5.2. This representation is built on top of low-level CNN features and is made of a large set of visual concept detectors learned with scalable linear classifiers. Each test image is compared to all detectors and the final Semfeat representation is composed of a handful of most salient concepts that are detected in the image. While in D5.2, detectors were used directly in order to derive image annotations, here we exploit Semfeat to first compute image similarity between the test image and a reference dataset and then analyze the annotations of neighbors in order to predict new annotations. Formally, if the Semfeat representation of a test image is written as:

$$S(I_t) = \{(C_1, w(C_1, I_t)), (C_2, w(C_2, I_t)), \dots, (C_N, w(C_N, I_t))\}$$

With C_j - the j[th] visual concept detector of the Semfeat representation and $w(C_j, I)$ – the probability of appearance of C_j given image I and N – the total number of visual concept detectors modeled in Semfeat.

Semfeat is sparse and, in practice only a few probabilities $w(C_k, I)$ will be non-null. Sparsity is an important property since, as we mentioned in D5.2, it enables a representation of the reference dataset using an inverted index structure which stores, for each visual concept, only the image-concept pairs with non-null probabilities. The inverted index can be written as:

$$C_1 = \{(I_x, w(C_1, I_x)), (I_y, w(C_1, I_y)), \dots, (I_z, w(C_1, I_z))\}$$
$$C_2 = \{(I_a, w(C_1, I_a)), (I_b, w(C_1, I_b)), \dots, (I_x, w(C_1, I_x))\}$$
$$\dots$$
$$C_N = \{(I_d, w(C_1, I_d)), (I_a, w(C_1, I_a)), \dots, (I_e, w(C_1, I_e))\}$$

The similarity of a test image I_t with the images of the reference dataset is computed by going through the list of concepts and summing-up scores of reference images that are associated with each non-null visual concept associated with I_t. Due to the sparsity of the Semfeat, this search operation is very fast since only a handful of the N modeled concepts are involved. We assume that only the top *k* neighbors (*k*=4 in the example hereafter) from the reference dataset are used to predict the annotations of image I_t, and that these neighbors are I_x, I_a, I_d and I_e, with user-provided annotations as follows:

$$A(I_x) = \{T_1, T_2, T_3\}$$

$$A(I_a) = \{T_1, \quad T_2, \ T_4\}$$
$$A(I_d) = \{T_1, \quad T_4, \ T_5\}$$
$$A(I_e) = \{T_1, \quad T_4, \ T_6\}$$

With T_i – user-contributed annotations (tags) associated with the reference images.

We devised two strategies to assign annotations to I_t, based either on simple counts of user-contributed annotations presented above or on user counts of the same annotations, noted as A_{sim} and A_{usr}. The hypothesis we make for the second strategy is that it will reduce the effect of bulk tagging, i.e. assignment of the same group of tags to a large set of images by a user, which is known to have a negative impact on data mining tasks with user-contributed data (O'Hare & Murdock, 2013). In our toy example, if we assume that I_x is contributed by a first user and I_a, I_d and I_e by a second user and that we want to retain only two terms in each case, the annotations will be:

- $A_{sim}\{I_t\} = \{T_1, \ T_4\}$ since these tags appear respectively four and three times in the user-contributed annotations
- $A_{usr}\{I_t\} = \{T_1, \ T_2\}$ since, even though T_4 has a higher total count than T_2, the last tag is contributed by two different users while T_4 is assigned by only one user.

2.2.2. Concept detection in presence of textual annotations

When user-contributed annotations are readily available, an obvious choice is to concatenate them and automatic annotations are obtained with concept detectors (D5.2) or with the similarity based method described in the preceding section. Beyond this simple approach, it is possible to combine existing annotations and image content representations in a more complex way and here we present a text-image fusion method called local soft tag coding. This representation relies on locality constrained coding (Yu et al., 2009), paired with max-pooling aggregation, which was successfully used to describe low-level image content, to represent user-contributed annotations. User-contributed textual annotations are lifted in a high-dimensional space using either ESA vectors (D5.1), Flickr co-occurrences or WordNet distance (Wu & Palmer, 1994). This operation is performed in order to provide a finer-grained textual representation that accounts for word relations in large-scale linguistic resources that encode complementary relations between words: Wikipedia includes encyclopedic knowledge, Flickr contains relations for a photographic language and WordNet is structured as a hierarchy.

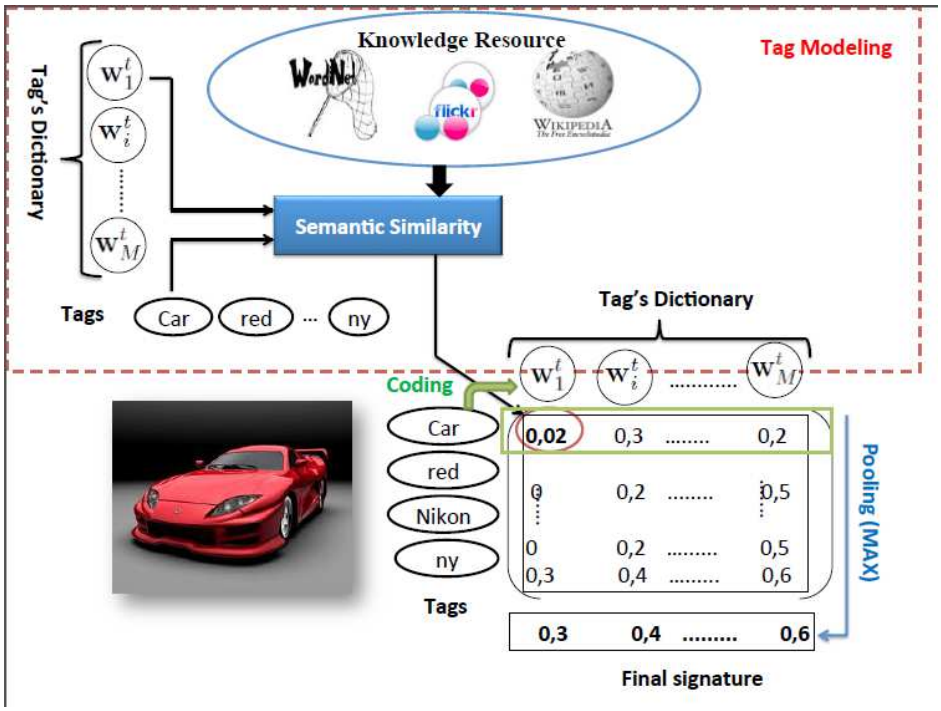

Figure 2. Illustration of local soft tag coding procedure. User-contributed annotations (car, red, ny) are mapped on a large tag dictionary in order to extract their similarity with the elements of this dictionary. The pooling step uses a MAX operator which selects the maximum value associated with dictionary elements (W_i^t). The final signature is a projection of the initial annotations on the tag dictionary.

Local soft tag coding is illustrated in Figure 1Figure 2 with an example of a scarcely annotated image.

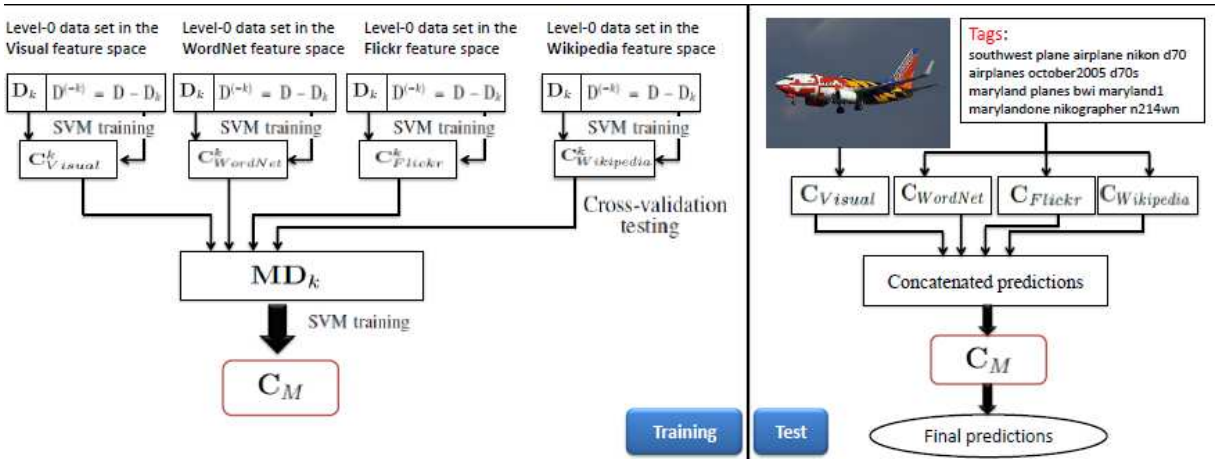

Figure 3. Illustration of stack generalization approach. The left side presents the stack generalization during the training phase, which includes a separate coding of visual and textual channels, followed by a late fusion using SVM based training in order to obtain multimedia features. The right side of the figure presents the test phase that compares the features of the test item to the pre-trained multimedia model in order to propose annotations. SVM classifiers are used to create models for the different concepts that need to be tested.

The three tag codings obtained with the three linguistic resources are then combined with visual representations which can be obtained with different low-level features, including bags of visual words or CNN features, using a stack generalization approach illustrated in Figure 3. Naturally, the training step is computationally complex and is thus performed offline, while the test step is performed live since it only involves a comparison of the test features with the

pre-trained multimedia concept models in order to predict the most probable concepts of the test image.

2.3. Evaluation and testing

2.3.1. Concept detection in absence of textual annotations

We implemented the visual similarity based concept detection using the best content based image retrieval configuration from Section 2.3 of D5.2. This configuration is based on Semfeat version built on top of 30,000 Flickr groups, with a large negative class composed of 100,000 images. To test our method to a state-of-the-art approach, we have compared it to the Clarifai annotation tool which implements the concept detection approach that achieved the best performance during the ImageNet LSVR Challenge[8]. Clarifai uses proprietary algorithms that leverage CNNs and claims to "recognize tens of thousands of categories, objects, and tags in any image". To facilitate reproducibility, the comparison is made on a subset of 1000 images of the publicly available MIR Flickr 25K dataset (Huiskes & Lew, 2008). Since our purpose is to reproduce user-contributed annotations, we use these last as ground truth and the purpose is to predict them automatically. Clarifai provides the top 10 annotations for each test image and, consequently, we can evaluate tagging precision up to this depth. We use the standard P@1, P@5 and P@10 levels which account for the number correct tags after 1, 5 and respectively 10 predictions.

Figure 4. Examples of annotation results. flickr anno stands for the user-contributed annotation that are used as ground truth for the evaluation ; CLARFAI and SEMFEAT are the top 10 (from left to right) tags obtained with the two systems compared here. For SEMFEAT, the A_{usr} strategy is used for tag aggregation. The tags presented in bold are those that match user-contributed annotations.

[8] http://www.image-net.org/challenges/LSVRC/2013/

Method	P@1 [%]	P@5 [%]	P@10 [%]
Clarifai	1.41	3.88	5.57
A_{sim}	1.91	5.23	7.42
A_{usr}	1.92	5.48	7.85

Table 1. Multimodal concept detection performance in absence of user-contributed annotations. P@X stand for the precision of results after X results, which is computed as the number of correctly predicted tags divided by the total number of tags from the ground truth. Results reported with A_{sim} and A_{usr} are given for k=2,000 neighbors, a value which was found to be optimal in preliminary tests.

The results obtained with Clarifai and the A_{sim} and A_{usr} methods described in Subsection 2.2.1 are presented in Table 1. It is noteworthy that the intersection between user-contributed and automatically predicted tags is rather low for all tested configurations. As illustrated in Figure 4, this situation is explained primarily explained by the fact that an image can be described with a large number of tags and the choices made by the users are not easy to reproduce. For instance, the user annotation of Figure 4 (d) is pudding while some of the tags found with the automatic methods (i.e. “food”, “bowl”, etc.) are also relevant. Equally important, user tags are often personal in nature and have no link with the image content from a social perspective. This is, for instance, the case for Figure 4 (c) where the tag “nena” probably refers to the name of the depicted cat.

When comparing the automatic methods tested, both versions of similarity based annotation proposed here clearly outperform Clarifai annotations at all precision levels. This result validates our approach compared to a state of the art large-scale concept detection and annotation system. The comparison of A_{sim} and A_{usr} is slightly favorable to the later method and this result indicates that the reduction of bulk upload effect is beneficial for the overall performance of the approach.

A further examination of the automatically predicted annotations shows that a wide majority of them are generic tags. Using a sample of 10 million Flickr metadata, we computed a list of most frequent 1000 tags and compared the annotations obtained with our method and with Clarifai with these tags. In both cases, over 98% of the automatically predicted tags were part of the 1000 most frequent tags list. The similarity based methods often fail to add specific concepts to the test image. For our approach, this shortcoming is explained by the fact that annotations are obtained by summing up the occurrences of tags over a rather large number of visual neighbors. This procedure naturally favors popular tags over more specific ones. In USEMP, one objective is to propose precise and specific insights about the content shared by the users and the visual similarity based method is probably less adapted than the D5.2 implementation of large-scale concept detection. This last method tests the content of the image against all the available models and is less likely to bias the results towards generic concepts.

2.3.2. Concept detection in presence of textual annotations

The evaluation of concept detection in presence of textual annotation is performed using the PascalVOC 2007 dataset (Everingham et al., 2010), which contains challenging images of 20 diversified concepts. We first test the performance of the textual representations and then combine them with bags of visual words and CNN features. Our results are compared to the multimodal fusion approach presented by (Guillaumin et al., 2010), which exploits hard coding of textual features (i.e. a word is either present or absent and is thus coded by 0 or 1) and a high dimensional visual content representation based on SIFT. Results are presented

using the Mean Average Precision (MAP) measure, with higher values standing for better performance.

Coding scheme	MAP
Hard coding (Guillaumin et al., 2010)	0.433
WordNet	0.494
Flickr	0.516
Wikipedia	0.513
WordNet+Flickr+Wikipedia	0.518

Table 2. Results on PascalVOC 2007 with different tag coding schemes.

The results presented in Table 2 illustrate the benefits of our tag coding schemes compared to the hard coding described in (Guillaumin et al., 2010). Results are improved with any of the three linguistic resources tested, with best results obtained with Flickr and Wikipedia. The combination of the three representations brings only marginal improvement compared to separate Flickr and Wikipedia representations.

Method	Visual	Textual	Multimodal
(Guillamin et al., 2010)	0.531	0.433	0.667
TXT + BOVW	0.521	0.518	0.683
TXT + Overfeat	0.696	0.518	0.780

Table 3. Results of PascalVOC 2007 with textual, visual and multimodal content representation. TXT is the combination of WordNet, Flickr and Wikipedia representations presented in Table 2. BOVW is the SIFT based bag of visual words used by our method. Overfeat is the CNN feature extracted from layer 18 of the Overfeat small network.

The results presented in Table 3 confirm that the combination of textual and visual features is beneficial for all presented approaches. This result is explained by the complementarity between these types of features. Both text-image combinations proposed here outperform the one introduced by (Guillaumin et al., 2010). Somewhat surprisingly, the combination of textual features and BOVW brings only 1.5% improvement while the text alone is 8.5% compared to the hard coding proposed by (Guillaumin et al., 2010). Confirming the results reported for the ImageNet dataset (Russakovsky et al., 2014) the CNN based features perform much better than any textual and visual individual features proposed before and even better than their combination. The fusion of textual and CNN features bring consequent improvement of overall results (8.4%) compared to the use of CNN features only.

2.4. Implementation and usage

USEMP scenarios are focused on extracting specific visual knowledge which is likely to be transformed into interesting insights for the user, especially when combined with any user-contributed textual annotations that might be associated with the image. In this first development step, the multimodal concept detection will operate a fusion between any annotations provided by the user and the automatically predicted annotations obtained from the visual concept detection method presented in Section 2 of D5.2. The more advanced late fusion of textual and image features that exploits textual resources (Wikipedia, Flickr, WordNet) poses scalability problems related to the high dimensionality of these text representations. Consequently, it will be provided at the end of the second iteration of this deliverable in order to be integrated in the final round of pilot tests.

The multimodal concept detection wrapper is implemented in Perl and can be called with the following command:

perl multimodal_fusion.pl [user-annotation-file] [auto-annotation-file] [mm-annotation-file]

The command and parameter files are further explained in Table 4.

Program	**Description**
multimodal_fusion.pl	Perl script that combines manual annotations (if any) and automatic annotations of an image.
File	**Description**
user-annotation-file	File that contains the user contributed annotations associated to an image (if available).
auto-annotation-file	File that contains the automatic annotations associated to an image.
mm-annotation-file	File that contains the combination of user contributed and automatic annotations associated to an image.

Table 4.Multimodal concept detection usage.

2.5. Next steps

The obtained results are interesting and future work is foreseen in two main directions, related to the scientific advancements of the task and to its integration in the USEMP framework. From a scientific point of view, we will focus on:

- Reducing the dimensionality of text representations for late fusion with visual features.
- Merging the results of multimedia fusion into the higher level privacy dimensions described in WP6. For instance, a detailed consumer profile could be obtained through a linkage between multimedia concepts and the IAB taxonomy[9].
- Combining other textual and visual insights gained through text and visual mining done in D5.1 and D5.2. An example of such fusion is that between product mentions in user texts and logos/product depictions from images.
- Merging the results of visual similarity-based and of learning-based annotation methods in order to have both generic and specific image descriptions.

From a USEMP integration perspective, multimodal concept detection will be included in the architecture after the pre-pilot and used during the tests performed toward the end of the second period as part of WP8.

[9] http://www.iab.net/QAGInitiative/overview/taxonomy (accessed on 26/12/2014)

3. Multimodal location detection

As we mentioned in D5.1 and D5.2, upstream work in WP4 and WP6 showed that location is one of the eight core privacy dimensions analyzed in USEMP. Multimodal location detection is a combination of location detection work done separately for textual and visual modalities. This module proposes a late fusion of results obtained with individual modalities in order to improve overall results. Following the work started in D5.1, we have notably investigated the use of confidence indices in order to decide which modality should be used for each multimedia item. Equally important, we exploited the information regarding the data source in order to propagate annotations through time and improve overall results.

3.1. Related work

Following early work on textual and visual modalities, presented respectively in (Serdyukov et al., 2009) and (Hays et al., 2008), different research groups have investigated the usefulness of combining the modalities in order to improve results. (Gallagher et al., 2009) were among the first to propose the use of multimodal information for content geolocation. They exploit the complementarity between tags and visual features to build location probability maps and show that the modality combination has a beneficial effect. (Kelm et al., 2011) use a hierarchical approach to model the geographical space and propose a late fusion of textual and visual features in order to improve geolocation. (Choi & Li, 2014) recently proposed a combination of textual and visual features in which the estimations are done with the visual modality whenever the textual one has low confidence. Visual similarities are computed using the Geo-Visual Ranking method, which relies on the use of SURF and color features. Their results are improved for small geolocation ranges (10 to 100 meters) but precision decreases for larger precision ranges. This finding could be explained by the fact that visual features capture mostly very local similarities (i.e. same point of view of a scene/POI) while tags capture larger range similarities (i.e. Notre Dame de Paris images are distributed in a radius around the actual location of the cathedral). Differently from visual features, textual annotations can capture geolocation information at different scale. For instance, “Notre Dame de Paris facade” links to highly localized information, “Paris, France” links to city level information, while “France” alone links to country level information. These different granularity textual annotations can be exploited in order to provide geolocation at different geographical scale. (Li et al., 2014) test the use of CNN features (Overfeat) for visual geolocation and obtain a slight improvement of results up to 1 km precision when combining visual and textual modalities.

3.2. Method description

Our fusion method is similar in spirit to the one proposed in (Choi & Li, 2014) since it exploits confidence scores of individual modalities in order to exploit either textual or visual based location predictions. Given that the textual modality gives substantially better results, it is used by default and the visual modality, based on adapted CNN features, is exploited only if the prediction score is over a confidence threshold. The visual modality exploits a variant of the GVR algorithm (Li et al., 2013) which performs a location based clustering of similar images in order to determine the most probable location. Given a set of *k* neighbors of an image, the top *t* among these neighbors are used as seed in order to count the number of similar images that are found within a radius *r* from the position of the seed. The most

probable location will be that of the seed which has the highest number of neighbors within the radius *r*. When combined with the textual predictions, a threshold on minimum size (s_{min}) of the largest spatial cluster is exploited. This value of this threshold is determined empirically using a sizeable validation set. The main difference with existing work comes from the visual features used. As we mentioned, we used the CNN models adapted for POI recognition that were introduced in Section 3 of D5.2. An important property of these algorithms compared to the features exploited in state-of-the-art approaches is that the dimensionality is significantly reduced. In our case, each image is described by 256 dimensions, versus 20,000 dimensions for SURF (Li et al., 2013) and 4096 dimensions for Overfeat (Li et al., 2014).

3.3. Evaluation and testing

The evaluation of the multimedia fusion method based on the protocol proposed for the MediaEval 2014 Placing Task (Choi et al., 2014). Two random subsets of 50,000 each of the 500,000 test images[10] were used for validation and testing. The training was done with 6,8 million images, including: (1) 5,000,000 training images that are sampled from the larger YFCC[11] dataset in order to provide a representative coverage of different world regions and (2) 1.8 supplementary images that depict POIs and are distinct from the test and validation sets. We have downloaded the test and training images and used the adapted CNN models (Section 3.2 of D5.2) to extract features from the full dataset. To speed-up execution, we computed PCA versions of the features and retained the first 256 dimensions of the PCA vectors for the experiments. This choice is motivated by our findings in D5.2, indicating that larger PCA vectors don't yield better performance while lower dimension representations are less accurate.

Features used	P@0.1 km	P@1 km	P@10 km
TXT	0.016	0.238	0.414
VIS	0.016	0.04	0.058
MM	0.021	0.241	0.415
Other best	0.043	0.222	0.39

Table 5. Geolocation prediction performance on the MediaEval 2014 Placing Task dataset. Accuracy is measured using precision at three granularity levels. P@X km is the proportion of test items placed at least than X kilometers from their true location. These precision ranges are chosen because they are the most likely to be useful in USEMP, where we are mainly interested in location detection up to city scale. "Other best" refers to the best multimodal system presented at MediaEval 2014 (Kordopatis-Zilos et al., 2014).

In Table 5, we present the results obtained with the following methods:

- TXT – text based location from D5.1, using only internal training data in order to improve comparability with state of the art approaches;
- VIS – visual based location prediction that exploits the PCA-compressed CNN models introduced in Section 3.2 of D5.2 test items and our version of the GVR algorithm (Choi & Li, 2014). This result is obtained with t=5 seeds, *k*=40 visual neighbors used for GVR clustering and a radius r=*1* km. As we mentioned, these values were selected using a validation set that is composed of 50,000 images.

[10] There are also 10,000 videos that are not considered in this evaluation since USEMP is focused only on still images.
[11] http://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67

- MM – multimodal approach that exploits confidence scores to combine TXT and VIS. the visual modality is used only when the size of the largest spatial cluster is $s_{min} \geq 20$ photos. This value was also determined on the validation set.

The presented results confirm those reported by (Kordopatis-Zilos et al., 2014), (Choi & Li, 2014) or (Li et al., 2014) and confirm that the visual location prediction still lags well behind its textual counterpart. For instance, in the VIS system, 4% of the photos were placed at less than 1 km from their true coordinates, a value that is slightly higher than the 3.88% reported by (Choi & Li, 2014) with the use of SURF features that have a higher dimensionality. Out of 50,000 images, there were only 1,033 that were geolocated with the visual modality since the rest of them had a spatial clustering score lower than the optimal threshold obtained on the validation set ($s_{min} < 20$). As a result, the score improvement for P@1km is only 0.3% for the MM method compared to the textual location prediction.

A qualitative analysis of the visual location detection allowed us to find three main issues that might affect its performance:

- Many geotagged images uploaded on Flickr do not have a visual location component. For instance, they can depict close-ups of faces, animals or plants and objects that can be found anywhere. In these cases, the visual based geolocation is impossible and the items can be localized only if they are accompanied by location related tags.
- Insufficient training data. Even though millions of images are used for training, they are clearly not sufficient in order to represent the very large number of scenes that can be depicted by geolocated images. In addition to the main experiment run with 6.8 million training images, we have made a test with the 5 million training images from the Placing Task dataset. P@1km is 4.1% in the first case and 3.8% in the second. While this difference is not very high, it indicates that adding more geolocated training data would be beneficial for the geolocation performance. When analyzing separately the 1033 images that had a spatial clustering score equal or higher than 20, P@1km is 67.2% and this result brings further support to the need for more training data.

The CNN model used was tuned for POIs (i.e. recognizable objects) and fails to geolocate correctly other types of geographically recognizable images, including city panoramas and natural scenes.

3.4. Implementation and usage

Multimodal location detection combines the results obtained with the textual and visual detection modules implemented as part of D5.1 and D5.2 respectively. It takes the results of these two modules as input and outputs a multimodal prediction as output. The wrapper is implemented in Perl and can be called with the following command:

perl multimodal_location.pl [txt-prediction-file] [vis-prediction-file] [threshold] [mm-prediction-file]

The command and parameter files are further explained in Table 6.

Program	**Description**
multimodal_location.pl	Perl script that combines textual and visual location predictions based on a confidence threshold.
File	**Description**
txt-prediction-file	File that contains the textual location detection results as provided

	by the module implemented as part of D5.1.
vis-prediction-file	File that contains the visual location detection results as provided by the module implemented as part of D5.2.
threshold	Confidence score of the visual location - used in order to select the modality for location detection.
mm-prediction-file	File that contains the combination of textual and visual location detection results.

Table 6. Multimodal location detection usage.

3.5. Next steps

From a scientific point of view, the multimodal location detection module is considered mature and its evolution will depend only on the developments of textual and visual detection techniques developed as part of D5.1 and D5.2. Notably, we would like to test what happens to the performance of visual location prediction if more training data are used and if the CNN model is better adapted to the dataset.

From a USEMP integration point of view, the multimodal location detection module is provided to the industrial partners for integration the USEMP system developed as part of WP7. It will be integrated an used during the first pilot tests.

4.Relevance- and Diversity-based Reranking

Despite the high accuracy that we achieved in the detection of private concepts with the concept detection models described in D5.2 (visual content mining), these models are still far from perfect. In addition to our efforts to improve the accuracy of these models via the use of better visual and textual descriptors and more sophisticated classification and fusion approaches, we hypothesize that the use of additional training examples through the collection of feedback from user interaction with the system could further enhance the accuracy of the privacy information extraction module described in D6.1. User feedback is especially important given the fact that we are not solely interested in correctly detecting the concepts depicted by users' images, but in a fine-grained distinction between those images that are perceived to be of private nature and those that are not. To this end, we developed a relevance- and diversity-based reranking method that has a two-fold goal: a) to improve the presentation of the results of the privacy scoring module to each particular user, b) to motivate users to provide relevance feedback since the more feedback they provide the better the privacy scoring is expected to become for both themselves and other users (Ferecatu et al., 2008).

Below, we describe in more detail how this reranking method fits into the USEMP system:

- The system will contain a component that given a private concept of interest to the user (e.g., drinking, smoking, etc.) will provide a ranking of the user's images (either images from his personal photo collection or those that have been uploaded on his OSN account) with respect to that concept.
- Following common practice in computer vision (Guillaumin et al., 2010), the most straightforward way to perform this ranking is based on the confidence scores predicted from the corresponding private concept detection model. These scores are chosen because they express the confidence of the algorithm in the automatic prediction (the higher the confidence is, the more likely it is that the algorithm prediction is correct). Since the model is not perfect, we expect irrelevant images within the top results.
- In order to improve the system's accuracy and to empower the user with the ability to "teach" a user-specific, customized definition of the private concept, the system could allow the user to explicitly designate through the UI which of the top images (typically users inspect only top results) are indeed relevant to the concept and which are not. In addition, the system could collect implicit relevance feedback by monitoring the user's interaction with the OSN (removal of images).
- The USEMP system can then build new privacy scoring models that are trained using both the existing training examples and the new examples (contributed through user feedback) in order to provide a better ranking of the images with respect to the concept. In this case, both user-specific and concept-specific (contributed by other users for the same private concept) examples could be utilized with weights that reflect their relative importance for the user (i.e. user-specific examples should receive higher weights).
- A problem with presenting a ranking of the images that is based exclusively on the classifier's confidence score is that many similar images could be present in top results. This has negative impact not only on the user's satisfaction (a user would prefer it if the system could present and ask feedback for a representative image from

a group of similar shots rather than overwhelming him/her with many similar images) but also on the quality of collected feedback since receiving relevance annotations for similar items is less useful for a model compared to receiving feedback for diverse items (Hoi et al., 2008), (Dagli et al., 2006). Therefore the system should ideally present a ranking where images that are relevant to the private concept but also different from the rest of relevant images are ranked higher.

Figure 1Figure 5 provides an example of how a relevance- and diversity-based reranking method could be used to enhance the initial ranking of a user's images for the private concept of "drinking". As we see, the initial ranking might promote at the top places images that are relevant to drinking but not of a private nature (im1-4). E.g. im3 depicts people (we assume that the user is included) drinking coffee during a conference. Although im3 is relevant to drinking, the user would probably not consider it private. Moreover, the initial ranking presents the highly similar images im1 and im2, both at the top of the ranking. This is a trivial result that would not be appreciated by the user. Besides, receiving user feedback for these two very similar items will not be as useful as receiving feedback for dissimilar images that are both ranked high by the private information extraction module. The second row of Figure 5 presents the optimal ranking that we would like to present to the user.

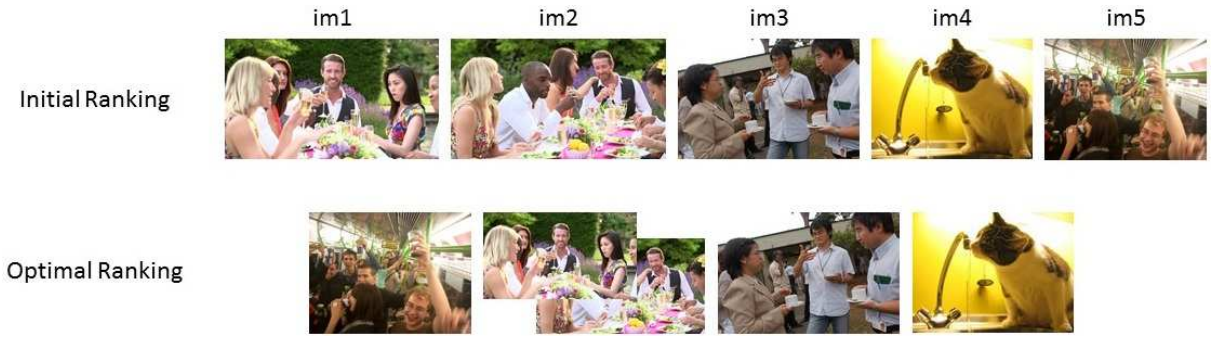

Figure 5 : Example showing the ranking of a user's images for the private concept of "drinking" produced by the private information extraction module (1st row) and an optimal ranking after applying a relevance- and diversity-based reranking method (2nd row).

The scenario described above bears many similarities to the topic of diverse retrieval where the task is to present a set of results that are at the same time relevant to the query but also exhibit diversity. Diversity in image retrieval was the focus of the "Retrieving Diverse Social Images" (RDSI) benchmarks of MediaEval 2013 (Ionescu et al., 2013) and 2014 (Ionescu et al., 2014) that adopted a landmark retrieval scenario. In the absence of a more appropriate evaluation setting (at the moment of writing the deliverable) for the reranking method that we developed for USEMP, we used the RDSI setting and dataset.

4.1. Related work

One of the first and seminal works on diversity in information retrieval is the work of (Carbonell & Goldstain, 1998). Recognizing that in the context of text retrieval and summarization, pure relevance ranking is not sufficient, they proposed Maximal Marginal Relevance (MMR). MMR is a reranking method that linearly combines independent measurements of relevance and diversity (their relative weight is a user-tunable parameter) into a single metric that is maximized in a greedy, iterative fashion. More recently, a similar formulation of the diversification problem was given by (Deselaers et al., 2009) 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 the optimization of a criterion that linearly combines relevance and diversity. However, (Deselaers et al., 2009) gives a more general formulation and uses dynamic programming algorithms to perform the optimization in addition to the greedy, iterative algorithm presented in (Carbonell & Goldstain, 1998). In our work, we adopt the formulation of (Deselaers et al., 2009) but combine it with a supervised definition of relevance that leads to significantly better performance. Also, compared to (Deselaers et al., 2009), where different modalities were combined in an ad-hoc way, the use of learning allows us to develop a more principled and effective way of combining multiple features.

Diversity in social image retrieval was the focus of the MediaEval 2013 and 2014 RDSI benchmarks that attracted the interest of many groups working in this area. Most participants developed diversification approaches that combined clustering with a strategy to select and return representative images from each cluster. Our MMR-based approach has the advantage of targeting the diversification problem in a more straightforward way compared to clustering-based approaches which first try to solve a different and presumably more difficult problem (i.e. finding groups of similar images). Also, despite the fact that most systems involved a mechanism to improve relevance before enforcing diversity, the majority did not exploit relevance annotations. Instead, top-performing solutions (Jain et al., 2013), (Dang-Nguyen et al., 2014) used specialized filters (e.g. face and blur detectors) and hand-coded rules (distant images are irrelevant) in order to discard irrelevant images according to the verbal definitions of relevance and irrelevance given by the task organizers. By learning the concept of relevance through the use of query and application-specific relevance annotations, our method can adapt automatically to different queries and retrieval scenarios and thus represents a more general solution.

4.2. Method

4.2.1. Problem definition

Let q be a query (private concept) and $I = \{im_1, \dots, im_N\}$ be a ranked list of user images that have been ordered according to the relevance scores provided be the USEMP system. Although the quality of the results depends on the specific concept and model, we expect that I will typically comprise both relevant and irrelevant images[12] and that some of the relevant images might contain duplicate information. The goal of our reranking method, is to refine the initial ranking of the images in I so that relevant images are ranked higher than irrelevant and top positions contain as little duplicate images as possible. Since users usually inspect only the top few results, reranking the whole list is not needed and we instead request a K-sized subset of images from I that are as relevant (to the concept) and as diverse (with each other) as possible. Among the many measures that have been proposed in order to quantify the above qualitative goal is, for instance, the subtopic or cluster recall at K ($CR@K$) (Zhai et al., 2003) that measures the percentage of different subtopics (subconcepts) retrieved in the first K results. Note that a perfect $CR@K$ requires all K results to be relevant.

4.2.2. Maximal Marginal Relevance

[12] Here, relevant means that the USEMP system presented the user with images that depict the concept.

MMR formulates the reranking problem described above as an optimization problem where one tries to maximize a linear combination of relevance and diversity, the so called “marginal relevance”. According to the formulation given in (Deselaers et al., 2009), the objective is to find the K-sized set $S \subset I$ that maximizes the following utility function:

$$argmax_{S \subset I, |S|=K}\, U(S|q) = w * R(S|q) + (1 - w) * D(S)$$

where $R(S|q)$ is a measure of the relevance of S to the query, $D(S)$ is a measure of the diversity in S, and w is a parameter that controls the relative importance of relevance and diversity. w can be either adjusted by the user or tuned to optimize a particular quantitative measure (e.g. $CR@K$).

4.2.3. Relevance

In the work of (Deselaers et al., 2009), the relevance of a set of images S was defined as $R(S|q) = \sum_{im_i \in S} R(im_i|q) = \sum_{im_i \in S} s(im_i, im_q)$ where $R(im_i|q)$ denotes the relevance of each individual image to the query and $s(im_i, im_q)$ is a normalized similarity measure (e.g. cosine) between visual or textual representations of im_i and im_q. A limitation of such an unsupervised definition is that similarity does not imply relevance as conceived by users. For instance, two images depicting a person sitting on her couch will receive a high similarity according to common image representations and similarity measures. However, the existence of an ashtray in one of the two pictures makes it more relevant to the private concept of “smoking” compared to the other. Motivated by that, we developed a supervised relevance scoring method that exploits relevance annotations in order to induce a more accurate definition of relevance. More specifically, for each query q, we build a probabilistic model $h_q: X \rightarrow [0,1]$ that takes a *n*-dimensional representation of the image $X = R^n$ as input and outputs the probability that the image is relevant to the query. These probabilistic outputs ($h_q(im_i)$) replace the unsupervised similarity measurements ($s(im_i, im_q)$). In order to train this model, we assume the existence of a set of training $D_q = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ of m training examples where $x_i \in X$ is the input vector and $y_i \in Y = \{0,1\}$ is the class value with $y_i = 1/0$ denoting a relevant/irrelevant example.

In Section 4.3, we perform experiments using various ways to compose this training set, adapted to the RDSI task. In particular, we evaluate models trained using only a limited amount of query-specific relevant examples, models trained on relevant and irrelevant application-specific examples (i.e. examples that are not query-specific but have been collected from similar queries) and finally models trained on a mixture of the two types of examples that are shown to obtain the best performance. In the usage scenario of USEMP described previously, query-specific examples could be mapped to user-specific examples and application-specific examples to concept-specific examples collected from multiple users.

4.2.4. A Multimodal Ensemble Classifier

When relevance annotations are used, a further advantage over unsupervised approaches is that the combination of multiple features can be incorporated into the learning process. Here, we present Multimodal Stacking (MMS), an ensemble classification algorithm that learns how to combine the outputs of multiple, independently trained classification models (each one using a different type of features) in order to make a better relevance prediction. The algorithm is inspired from stacked generalization (Wolpert, 1992), a method for the fusion of heterogeneous classifiers, widely known as stacking. The training of MMS consists of the

following steps: Initially, k independent probabilistic classifiers $h_{qi}: X \rightarrow [0,1], i = \{1, \dots, k\}$ are built, one for each multi-dimensional feature representation $X_i \in R^{ni}, n_i > 1$. Each of these single-feature classifiers is then used to predict the classes of all training examples and their predictions are gathered to form a meta training set $D'_q = \{(x'_1, y_1), \dots, (x'_m, y_m)\}$, where the input vectors $x'_i = [h_{qi}(x_1), \dots, h_{qk}(x_k)]$ consist of the outputs of the single-feature classifiers. This meta training set is used to train a meta classifier $h'_q: X \rightarrow [0,1]$, where $X' \in R^k$ is the meta input space and its output is the probability that the image is relevant. At prediction time, the single-feature classifiers are first applied to classify the unknown instance and their outputs are used to form a meta-instance that is fed to the meta classifier which makes the final prediction. Compared to early fusion approaches, MMS has the advantage that features of different dimensionalities can be combined since all models contribute a one-dimensional feature to the meta classifier. Furthermore, additional one-dimensional features can be easily incorporated into the final model by directly augmenting the input space of the meta classifier.

4.2.5. Diversity

Assuming a ranking $im_{r1}, \dots, im_{rK}$ of the images in S, Deselaers et al. (2009) define diversity as $D(S) = \sum_{i=1}^{K} \frac{1}{i} \sum_{j=1}^{i} d(im_{ri}, im_{rj})$, where $d(im_{ri}, im_{rj})$ is the dissimilarity between the images ranked at positions i and j. Thus, high diversity scores are given to image sets with a high average dissimilarity. We notice that with this definition of diversity, an image set that contains pairs of highly similar (and therefore not diverse) images is allowed to receive a high diversity score if the average dissimilarity is high. This results in a direct negative impact on diversification measures such as $CR@K$. Therefore, we adopt the following more strict definition: $D(s) = \min_{im_i, im_j \in S, i \neq j} d(im_i, im_j)$ where the diversity of a set S is defined as the dissimilarity between the most similar pair of images in S.

4.2.6. Optimization

An exhaustive optimization of U has high complexity as it would require computing the utility of all $\frac{N!}{K!(N-K)!}$ K-subsets of I. Therefore, we apply a greedy, iterative optimization algorithm that was also used in (Carbonell & Goldstein, 1998), with appropriate changes to reflect our new definitions for relevance and diversity. This algorithm starts with an empty set S and sequentially expands it by adding at each step $J = 1, \dots, K$ the image im^* that scores highest (among the unselected images), to the following incremental utility function:

$$U(im^*|q) = w * h_q(im^*) + (1 - w) * \min_{im_j \in S^{J-1}} d(im^*, im_j)$$

where S^{J-1} represents S at step $J - 1$

4.3. Experiments

In our empirical evaluation, we adopt the application scenario of the 2014 Retrieving Diverse Social Images (RDSI) task of MediaEval (Ionescu et al., 2014) and follow exactly the same evaluation protocol in order to obtain comparable results. The task addressed the problem of result diversification in social photo retrieval. The RDSI participating teams were provided with an ordered list of up to 300 images returned by Flickr in response to a textual query for a specific Point of Interest (POI) and were asked to refine this list by providing a ranked list of up to 50 images that are both relevant and diverse representations of the query. Explicit

definitions were provided for both relevance (e.g. artistically deformed photos are relevant while photos that present an aspect of a POI that is not socially recognizable aren't) and diversity (e.g. different times of the day/year). The refinement and diversification process could be based on the information provided for each POI (Wikipedia page, up to five representative photos from Wikipedia, GPS coordinates), the metadata of the retrieved images (e.g. title, description, tags, GPS coordinates, etc.) as well as their visual content. During the task, participants were provided with an annotated development set of 30 queries (ground truth) - in order to build their approaches - as well as a test set of 123 queries - upon which they were evaluated - whose ground truth was disclosed only after the end of the task. Ground truth consisted of relevance and diversity annotations provided by experts for all images of each POI. Specifically, each image was first labelled as either relevant or irrelevant and then visually similar relevant images were grouped together into clusters. Performance on each query was assessed using the F1@20 metric that is equal to the harmonic mean of CR@20 and P@20 (the percentage of relevant images in the top 20).

We used three categories of features to represent the dataset: a) visual - computed directly from the image content; b) textual - computed from textual annotations, and c) meta - the metadata associated with the images. For all types of multi-dimensional features unit length normalization is applied and cosine similarity/distance is used for relevance/diversity computations.

Visual: After initial experiments with the features made available by the task organizers (Ionescu et al., 2013), we extracted the following state-of-the-art features that lead to significantly better performance:

VLAD: d=24,576-dimensional VLAD+CSURF vectors (Spyromitros-Xioufis et al., 2014) are computed using a 128-dimensional visual vocabulary and then projected to d' dimensions with PCA and whitening. Using d'=128 leads to near-optimal results for both relevance and diversity.

CNN: were described as part of D5.2, Section 3 and are based on an adaptation of a CNN architecture with POI-related data. Similar to VLAD, their size is reduced from 4096 initially to 128 dimensions in order to accelerate the retrieval process.

Textual: To generate textual features we first transformed each query and each Flickr image into a text document. For queries, we used a parsed version of the corresponding Wikipedia page and for Flickr images we used a concatenation of the words in their titles, descriptions and tags. Bag-of-words features (*BOW*) were then computed for each document using the 10K most frequent terms of the collection as the dictionary and term frequencies as term weights. We found that by repeating the terms in the image titles and descriptions two and three times respectively to increase their contribution in the similarity compared to the terms in the tags that are usually noisier, led to increased performance.

Meta: The following one-dimensional features were computed from the textual metadata and used as additional features in the meta input space of the MMS algorithm: distance from the POI, Flickr rank, number of views.

4.3.1. Supervised vs Unsupervised Relevance

In this section we compare the unsupervised variant of the MMR method (Deselaers et al., 2009) (*uMMR*), with the proposed supervised variant, named *sMMR*. We show experiments

when either a visual (VLAD)[13] or a textual (BOW) representation is used for both relevance and diversity to highlight potential differences between visual and textual features. The sMMR method can be instantiated with any classification algorithm. We choose L2-regularized Logistic Regression as it provided a good trade-off between efficiency and accuracy compared to other state-of-the-art classifiers in preliminary experiments. Depending on how the training set was composed for each query, three different variants of the sMMR method were created:

$sMMR_q$: The training set contains only the Wikipedia images or the textual representation of the Wikipedia page. Since these are positive examples, we add a few randomly chosen Flickr images from other queries as negative examples. $sMMR_q$ attempts to capture the query-specific notion of relevance.

$sMMR_a$: The training set is composed of Flickr images from other queries as positive and negative examples. $sMMR_a$ attempts to capture the application-specific notion of relevance.

$sMMR_{aq}$: The training set is composed of Flickr images from other queries as positive and negative examples as well as the Wikipedia page/images as positive examples. We found that simply combining the few query-specific positive examples with a significantly larger number of application-specific positive examples generates very similar models with the $sMMR_a$ variant. Therefore, we experimented with assigning higher weights to the query-specific positive examples in order to increase their contribution to the formation of the classification boundary. $sMMR_{aq}$ attempts to capture both the query and the application-specific notion of relevance.

Table 7 presents the test set Area Under Curve[14] (AUC) and F1@20 scores of uMMR and the variants of sMMR presented above. AUC is calculated on the relevance rankings produced by each variant (without diversification). When VLAD features are used, we observe that all sMMR variants outperform uMMR in terms of AUC. Interestingly, despite the absence of any query-specific relevance information, the sMMR $_a$ variant obtains a significantly better AUC than both uMMR and sMMR $_q$, suggesting that capturing an application-specific notion of relevance is important in relevance scoring using visual features. As expected, when the query and application-specific relevance information are combined in the sMMR $_{aq}$ variant, AUC performance improves even further, especially when a large weight is assigned to query-specific examples. With respect to F1@20, we notice that sMMR $_{aq*1000}$, the best performing variant in terms of AUC, also obtains the best F1@20 score (0.561) that is about 5% better than the 0.536 score obtained by uMMR.

When textual features are used, the uMMR method (the two variants coincide in this case) is only slightly outperformed by sMMR $_q$ and sMMR $_{aq*1000}$ in terms of AUC. Contrarily to when visual features are used, sMMR $_a$ obtains a significantly lower AUC compared to both uMMR and the sMMR variants that use query-specific information. This suggests that application-specific information is not sufficient to produce a good relevance scoring when this scoring is based only on the textual modality. Nevertheless, the relevance scoring produced by sMMR $_a$ is better than random scoring (AUC=0.5) and the sMMR $_{aq*1000}$ variant that

[13] Similar results were obtained with CNN features.
[14] AUC is a measure of the ranking accuracy, i.e. quantifies the extent to which relevant images are ranked above irrelevant ones.

combines query and application-specific information is again the best performer. With respect to F1@20, we observe that, with the exception of the sMMR $_{a}$ and sMMR $_{aq*1}$ variants, all other variants produce better F1@20 scores than uMMR. When sMMR $_{aq*1000}$ is used, F1@20 increases from 0.503 to 0.555, a 10% improvement.

Method	VLAD		BOW	
	AUC	F1@20	AUC	F1@20
uMMR	0.622	0.536	0.660	0.503
sMMR $_{q}$	0.624	0.535	0.666	0.553
sMMR $_{a}$	0.664	0.536	0.571	0.481
sMMR $_{aq*1}$	0.665	0.536	0.575	0.487
sMMR $_{aq*1000}$	0.693	0.561	0.670	0.555

Table 7: AUC and F1@20 performance (averaged over test set queries) of uMMR and sMMR using visual and textual features.

4.3.2. Multimodal Fusion Experiments

In the previous experiment, we evaluated instantiations of the sMMR method that used only a single type of features for relevance and diversity scoring. In this section, we want to evaluate the performance of the method when more features are combined to assign relevance scores. To combine multiple features for relevance scoring in sMMR, we use the MMS ensemble method presented in Section 4.2.4. More precisely, the sMMR $_{aq*1000}$ configuration is used for the single-feature models since it was found superior to other configurations in and L2-regularized Logistic Regression is used as classification algorithm for both the single-feature models and the meta model. The resulting method is called sMMR-MMS.

Table 8 presents the F1@20 scores obtained with sMMR-MMS using pairs of the three multi-dimensional features presented above. We also present results for sMMR-MMS when the three one-dimensional meta features are used within the MMS algorithm as described in Section 4.2.4. Results using each of the three features alone for both relevance and diversity are also reported to facilitate comparison. In all multi-feature instantiations, VLAD features were used for diversification as they gave the best F1@20 scores when tested with artificial classifiers of various AUC performances.

Features	**F1@20**	**Features**	**F1@20**	**Features**	**F1@20**
VLAD	0.561	VLAD+CNN	0.594	VLAD+CNN+meta	0.619
CNN	**0.572**	VLAD+BOW	0.578	VLAD+BOW+meta	0.590
BOW	0.555	CNN+BOW	**0.608**	CNN+BOW+meta	**0.631**

Table 8: F1@20 performance of sMMR-MMS using various combinations of features.

Looking at the results obtained with single-feature instantiations we observe that the performance obtained with CNN features is significantly better compared to VLAD and BOW. With respect to two-feature instantiations we observe that F1@20 is always better than using any of the two features alone suggesting that the MMS algorithm can effectively fuse different modalities. As expected, larger improvements are obtained when a textual and a visual modality are combined (VLAD+BOW, CNN+BOW) compared to combining modalities of the same type (VLAD and CNN). Finally, we see that when the meta features are used, performance improves further in all cases. In particular, the CNN+BOW+meta combination obtains the highest F1@20 score followed by VLAD+CNN+meta and VLAD+BOW+meta. Compared to the best performing systems submitted at RDSI 2013 (Jain et al., 2013) and

2014 (Dang-Nguyen et al., 2014) our sMMR-MMS method with CNN+BOW+meta features obtains a 7.5% and 5.7% better performance respectively.

4.4. Implementation and usage

In terms of implementation we provide an executable jar file (reranking.jar) that can be used to perform the following operations on the RDSI task dataset:

1. Compute and store distances/similarities between all pairs of images in a RDSI location. To perform this operation use the following command:
   java –jar reranking.jar dc [data-folder] [feature-type]
   where
   [data-folder] full path to the root of the RDSI task dataset
   [feature-type] type of features (vlad/cnn/bow)

2. Create ARFF formatted training and test datasets that can be used for relevance learning. To perform this operation use the following command:
   java –jar reranking.jar md [data-folder] [feature-type]
   where [data-folder] and [feature-type] have the same meaning as above.

3. Build a relevance detection model on the training set and evaluate it on the test set. Also generates a relevance ordered file in the RDSI submission format. To perform this operation use the following command:
   java –jar reranking.jar eval [dataset] [smmr-variant] [submission-file]
   where
   [dataset] full path and filestem of the relevance learning dataset
   [smmr-variant] the variant of the sMMR method to evaluate (a/q/aq)
   [submission-file] full path to the RDSI submission file that will be generated

4. Performs diversity-based reranking on a relevance-ordered RDSI submission file. To perform this operation use the following command:
   java –jar reranking.jar pp [data-folder] [feature-type] [submission-file] [w]
   where
   [data-folder] full path to the devset or testset folder of ther RDSI task dataset
   [feature-type] type of features (vlad/cnn/bow) to use for diversification
   [w] w parameter of the reranking method

In addition to the jar file we also provide the part of the RDSI task dataset that is necessary for the computations.

4.5. Next Steps

The reranking method as well as the multimodal fusion algorithm that we developed obtained encouraging results in a diverse image retrieval benchmark. In the future, we would like to test our approach using an evaluation setting and dataset that are targeted to the USEMP usage scenario described in the beginning of the section. This involves the development of a tool that presents a ranking of a user's images with respect to a private concept and collects user feedback with respect to the presented rankings. This tool will be used to evaluate the effectiveness of the diverse reranking method in terms of both user satisfaction and feedback

quality, quantified as the amount of improvement in the accuracy of the privacy scoring models.

5.Conclusions and future work

During the first iteration of the project, work on developing multimedia mining modules was conducted in three main directions: concept detection, location detection and improving the feedback given to the end-users. After introducing concept detection based on learning and similarity based, we focused on a late fusion of textual and visual cues and showed that this combination has beneficial effect on the overall quality of results. An important challenge that we identified is to propose automatic annotations that include both generic and specific concepts and one thread of future work will consider this problem. Multimedia location detection was a second important work thread and was performed with a late fusion of textual and visual cues that exploits the confidence of visual content based location in order to decide which modality should be used. While only limited improvement is obtained when exploiting visual cues in addition to textual ones, the visual component is still very valuable whenever there are no textual annotations associated to images. A shortcoming of our experiments comes from the fact that the very large search space represented by all locations in the world is not well sampled by the reference dataset used for geolocation. Future experiments will examine the role of supplementary reference images. The third work direction explored here was the exploitation of user feedback in order to rerank the results obtained with automatic image mining methods. The obtained results are very interesting and future work will focus on the application of the method for privacy-related datasets.

In parallel to multimedia mining modules improvement, integration work will be carried out to make all modules available in the USEMP system. The modules are provided along with this report and will be progressively integrated on the platform before the end of the second reporting period (September 2015).

6.References

Carbonell, J., & Goldstein, J. (1998). 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 (pp. 335-336). ACM.

J. Choi, X. Li. (2014). The 2014 ICSI/TU Delft Location Estimation System. Working notes of MediaEval Placing Task 2014

J. Choi, B. Thomee, G. Friedland, L. Cao, K. Ni, D. Borth, B. Elizalde, L. Gottlieb, C. Carrano, R. Pearce, D. Poland. (2014). The placing task: A large-scale geo-estimation challenge for social-media videos and images. Proceedings of the 3rd ACM International Workshop on Geotagging and Its Applications in ACM Multimedia, 2014

Dagli, C. K., Rajaram, S., & Huang, T. S. (2006). Leveraging active learning for relevance feedback using an information theoretic diversity measure. In Image and Video Retrieval (pp. 123-132). Springer Berlin Heidelberg.

Dang-Nguyen, D. T., Piras, L., Giacinto, G., Boato, G., & De Natale, F. G. B. (2014). Retrieval of Diverse Images by Pre-filtering and Hierarchical Clustering. In MediaEval 2014 Workshop, Barcelona, Spain.

Deselaers, T., Gass, T., Dreuw, P., & Ney, H. (2009). Jointly optimising relevance and diversity in image retrieval. In Proceedings of the ACM international conference on image and video retrieval (p. 39). ACM.

M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman (2010) The Pascal visual object classes (VOC) challenge. International journal of computer vision 88 (2), 303-338

M. Ferecatu, N. Boujemaa, M. Crucianu (2008) Semantic interactive image retrieval combining visual and conceptual content description. Multimedia Syst. 13(5-6): 309-322.

A Gallagher, D Joshi, J Yu, J Luo (2009) Geo-location inference from image content and user tags. Computer Vision and Pattern Recognition Workshops, 2009

A.L. Ginsca, A. Popescu, H. Le Borgne, N. Ballas, P. Vo, I. Kanellos (2015) Large-Scale Image Mining with Flickr Groups. Proc of MultiMedia Modeling 2015, pp. 318-334

M. Guillaumin, J. Verbeek, and C. Schmid, (2010) Multimodal semisupervised learning for image classification," in IEEE CVPR, 2010, pp. 902 – 909.

Hoi, S. C., Jin, R., Zhu, J., & Lyu, M. R. (2008). Semi-supervised SVM batch mode active learning for image retrieval. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on (pp. 1-7). IEEE.

M.J. Huiskes, M.S. Lew (2008) The MIR flickr retrieval evaluation. Proc. of ACM ICMR 2008.

Ionescu, B., Menéndez, M., Müller, H., & Popescu, A. (2013). Retrieving diverse social images at mediaeval 2013: Objectives, dataset and evaluation. In MediaEval 2013 Workshop, Barcelona, Spain.

Ionescu, B., Popescu, A., Lupu, M., Gınsca, A. L., & Müller, H. (2014). Retrieving diverse social images at mediaeval 2014: Challenge, dataset and evaluation. In MediaEval 2014 Workshop, Barcelona, Spain.

Jain, N., Hare, J., Samangooei, S., Preston, J., Davies, J., Dupplaw, D., & Lewis, P. H. (2013). Experiments in diversifying Flickr result sets. In MediaEval 2013 Workshop, Barcelona, Spain.

P. Kelm, S. Schmiedeke, and T. Sikora (2011) A hierarchical, multimodal approach for placing videos on the map using millionsof flickr photographs. Proc. of SBNMA '11 , New York, NY, USA, 2011, pp. 15–20, ACM

G. Kordopatis-Zilos, G. Orfanidis, S. Papadopoulos, Y. Kompatsiaris (2014) SocialSensor at MediaEval Placing Task 2014. Working notes of MediaEval 2014.

X. Li, C.G.M. Snoek, M. Worring (2009) Learning social tag relevance by neighbor voting. Multimedia, IEEE Transactions on 11 (7), 1310-1322

Lin Tzy Li, Otávio Augusto Bizetto Penatti, Jurandy Almeida, Giovani Chiachia, Rodrigo Tripodi Calumby, Pedro Ribeiro Mendes Júnior, Daniel Carlos Guimarães Pedronette, Ricardo da Silva Torres (2014) Multimedia Geocoding: The RECOD 2014 Approach. Working notes of Mediaeval 2014.

N. O'Hare, V. Murdock (2013) Modeling locations with social media. Information Retrieval, 2013.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei. (2014). ImageNet Large Scale Visual Recognition Challenge. arXiv technical report: http://arxiv.org/abs/1409.0575

Spyromitros-Xioufis, E., Papadopoulos, S., Kompatsiaris, I., Tsoumakas, G., & Vlahavas, I. (2014). A Comprehensive Study over VLAD and Product Quantization in Large-Scale Image Retrieval.

Wolpert, D. H. (1992). Stacked generalization. Neural networks, 5(2), 241-259.

Z. Wu and M. Palmer (1994) Verb semantics and lexical selection," in Annual Meeting of the Association for Computational Linguistics, 1994, pp. 133–138.

N. C. Yang, C. M. Kuo, W. H. Chang (2013) Content-Based Image Feature Description and Retrieval. In Search Algorithms for Engineering Optimization, Dr. Taufik Abrão (Ed.), ISBN: 978-953-51-0983-9, InTech, 2013.

K. Yu, T. Zhang, and Y. Gong (2009) Nonlinear learning using local coordinate coding, Advances in NIPS, pp. 2223–2231, 2009.

Zhai, C. X., Cohen, W. W., & Lafferty, J. (2003). Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 10-17). ACM.

A. Znaidia, A. Shabou, A. Popescu, H. Le Borgne, C. Hudelot (2012) Multimodal feature generation framework for semantic image classification. Proc of. ACM ICMR 2012.