Multimed Tools Appl (2011) 55:27–52
DOI 10.1007/s11042-010-0584-1

Reliability and effectiveness of clickthrough data for automatic image annotation

Theodora Tsikrika · Christos Diou · Arjen P. de Vries · Anastasios Delopoulos

Published online: 17 August 2010

Abstract Automatic image annotation using supervised learning is performed by concept classifiers trained on labelled example images. This work proposes the use of clickthrough data collected from search logs as a source for the automatic generation of concept training data, thus avoiding the expensive manual annotation effort. We investigate and evaluate this approach using a collection of 97,628 photographic images. The results indicate that the contribution of search log based training data is positive despite their inherent noise; in particular, the combination of manual and automatically generated training data outperforms the use of manual data alone. It is therefore possible to use clickthrough data to perform large-scale image annotation with little manual annotation effort or, depending on performance, using only the automatically generated training data. An extensive

T. Tsikrika (✉) · A. P. de Vries
Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands
e-mail: theodora.tsikrika@acm.org

A. P. de Vries
e-mail: arjen@acm.org

C. Diou · A. Delopoulos
Multimedia Understanding Group, Electrical and Computer Engineering Department,
Aristotle University of Thessaloniki, Thessaloniki, Greece

C. Diou
e-mail: diou@mug.ee.auth.gr

A. Delopoulos
e-mail: adelo@eng.auth.gr

C. Diou · A. Delopoulos
Informatics and Telematics Institute, Centre for Research and Technology Hellas,
Thessaloniki, Greece

A. P. de Vries
Delft University of Technology, Delft, The Netherlands

presentation of the experimental results and the accompanying data can be accessed at http://olympus.ee.auth.gr/~diou/civr2009/.

Keywords Image annotation · Concepts · Supervised learning · Search logs · Clickthrough data · Collective knowledge · Implicit feedback · Reliability

1 Introduction

The application of supervised machine learning approaches in the automatic annotation of images and videos with semantic concepts requires the availability of labelled samples to be used as training data. Such annotated samples are typically generated manually, a laborious and expensive endeavour. Even though collaborative large-scale annotation efforts have been organised, e.g., in the context of the TRECVID[1] evaluation benchmark [2], the bottleneck still remains, given, in particular, the large number of semantic concepts estimated to be desirable in order to achieve higher retrieval effectiveness than the current state-of-the-art [11]. The situation is further exacerbated by the poor generalisation of concept classifiers to domains other than their training domain [39]; this implies that for achieving effective annotation, individual content owners need to carry out their own manual annotation exercise, a continual task for the many collections that keep expanding over time with new data.

To mitigate the high cost of enlisting dedicated annotators to manually label training samples, the use of alternative data sources acquired as a by-product of human activities on the Web has been recently advocated for training concept classifiers. Such data sources include the labelled images produced as a side effect of people playing enjoyable Web-based 'games with a purpose' [36], such as the ESP game [35], aka the Google Image Labeler,[2] and its variations [14, 37], or completing CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart) [34] based on images [8, 22] rather than text [38]. These ingenious approaches have generated large image datasets annotated with meaningful and accurate labels by people enticed to participate in games based on their desire to be entertained or required to verify that they are human users as a security measure before accessing specific Web services. Additional data sources include the user-generated multimedia content annotated with user-defined tags found in abundance in social media systems like YouTube[3] and Flickr.[4] In this paradigm shift, Web communities unknowingly share in the generation of large amounts of labelled data, which can then be used as annotated training samples for building concept classifiers [5, 6, 19, 26, 31–33].

The work presented in this paper also follows this research direction and is concerned with building concept classifiers that use automatically acquired labelled samples as training data. Focussing on the specific case of image annotation, it

[1] http://www-nlpir.nist.gov/projects/trecvid/

[2] http://images.google.com/imagelabeler/

[3] http://www.youtube.com/

[4] http://www.flickr.com/

proposes and investigates the use of a different (and largely untapped) source for obtaining such examples: the *clickthrough data* logged by retrieval systems. These data consist of the queries submitted by the users of such systems, together with the images in the retrieval results that these users selected to click on in response to their queries. This information can be viewed as a type of users' *implicit feedback* [18] that provides a "weak" indication of the relevance of the image to the query for which it was clicked on [7]. We refine the notion of relevance in this assumption by considering that the queries for which an image was clicked provide in essence a "weak" description (or *annotation*) of the image's visual content. Our aim, therefore, is to investigate whether images with such search log based annotations can serve as labelled samples in a supervised machine learning framework for training effective concept classifiers.

The primary advantage of using such annotated samples is that these are generated without any explicit user intervention and without any major effort on the part of content owners, since clickthrough data are gathered unobtrusively and in large quantities in search logs during the users' search-related interactions. Furthermore, most content owners are able to collect their own search logs and therefore produce training data (and associated classifiers) that are adapted to their collections, rather than having to rely on the use of external tagged sources and deal with cross-domain applicability issues.

On the other hand, the major shortcomings are that automatically acquired labelled data are sparse (they only cover the part of the collection that has been previously accessed) [7] and potentially noisy [28]. Manual annotations are reliable and based on clear visual criteria pertaining to the samples' visual content, whereas user-defined tags and logged queries tend to describe not only the visual content, but also the context of multimedia resources. This has been recently illustrated in an analysis showing that Flickr's users annotate their photos with respect to both their content and their context by using a wide spectrum of semantic tags [27]. Nevertheless, the use of large amounts of "noisily labelled" data might be the key in dealing with this quality gap. In particular, clickthrough data (and also tags assigned in a collaborative manner) could be considered as having further noise reduction properties, given that they encode the collective knowledge of multiple past users, rather than the subjective assessment (or tag assignment) of a single person.

This work examines the usefulness of clickthrough data, as alternative or complementary data sources to manual annotations, for training concept classifiers in the particular application of automatic image annotation. This paper builds on our previous work [30] that investigated the effectiveness of the proposed approach. This extension presents further analysis and insights on the results of our experimental evaluation, with particular focus on the reliability of the search log based training samples and the effect of the levels of noise in these samples on the precision of the concept classifiers.

The remainder of this paper is organised as follows. Section 2 discusses related work on the use of (i) data sources other than manually labelled samples as training data in the annotation of multimedia content, and (ii) clickthrough data in multimedia retrieval applications. Section 3 describes our approach, while Sections 4 and 5 present the set up and the results of our experiments. Section 6 concludes this paper by summarising our main contributions and findings and by outlining future research directions.

2 Related work

Research on using data sources other than manually labelled samples for training concept classifiers has thus far focussed on the use of publicly available tagged multimedia resources, particularly images from Flickr [5, 6, 19, 26] and videos from YouTube [31–33]. Positive samples typically correspond to tagged resources downloaded through these sites' search services in response to a text query corresponding to the concept name [5, 6, 26, 32, 33], with the possible addition of manually selected keywords [31]. In all these cases, negative samples have been randomly selected. To reduce the potential noise, false positives can be removed (i) manually, by restricting the initial search to categories deemed relevant to the concept in question [31] or by disambiguating the tags [26], and (ii) automatically, by eliminating resources with low visual similarity to manually annotated data [5]. False negatives can also be filtered out by removing resources tagged with the concept's name and its synonyms from the randomly selected negative samples [19]. Concept classifiers trained with Flickr images have been applied both to test sets consisting of other Flickr images [6, 26] and across domain to test sets comprising images from the PASCAL VOC Challenge[5] [19] or TRECVID videos [5, 26]. Similarly, concept classifiers trained with YouTube videos have been applied both to test sets consisting of other YouTube videos [32, 33] and across domain to TRECVID videos [31] and German TV news [32, 33].

These studies indicate that (i) building concept classifiers trained on tagged resources obtained from the Web is feasible, (ii) such classifiers achieve "fair" effectiveness when applied on data obtained from the same domain (i.e., other tagged resources) [6, 33], and (iii) such classifiers are outperformed in cross-domain settings by classifiers trained on manually annotated data obtained from the same domain as the target data, with the former though working well for some concepts [19, 31]. This latter observation reveals that much further research is needed in order to reach reliable conclusions on the usefulness of such resources as training data, since it is not clear yet whether this relative loss in effectiveness is due to noisy and unreliable labels or due to domain differences. Overall, learning from Web resources is an attractive but currently unfulfilled research direction for the task of automatic annotation of multimedia content with semantic concepts.

Labelled samples can also be automatically acquired by processing the clickthrough data logged by retrieval systems. Such data have been used in several Information Retrieval (IR) applications, e.g., for generating surrogate document representations [24], for query suggestion [3], and, most importantly, as training samples for learning [15] or tuning [21] retrieval functions. Such approaches are being developed in the context of the recently resurgent and increasingly active research field of 'learning to rank' [17] that investigates the application of machine learning to IR. In multimedia retrieval applications, the use of clickthrough data has been far more limited, most probably due to the lack of publicly available search logs from multimedia search engines. Clickthrough data have been used in multimedia settings for labelling images with the terms in the queries for which these images or the Web pages containing them have been previously clicked [1] and also for probabilistically ranking images with respect to a given textual query based on a Markov random walk

[5] http://www.pascal-network.org/challenges/VOC/

model applied to the *clickgraph*, i.e., the bipartite graph where one set of vertices corresponds to queries and the other to images, and an edge denotes that an image has been clicked for a query [7]. To the best of our knowledge, though, there has been no previous work on incorporating the clickthrough data into the concept learning process.

3 Our approach

This section describes our approach for selecting 'training images' based on clickthrough data; such images can then be employed as labelled samples for training concept classifiers in automatic image annotation. In the following, we assume the existence of a collection $\mathbf{I}$ of images and of pairs of the form $(\mathbf{q}, \mathbf{I_q})$, where $\mathbf{q}$ is a text query (possibly consisting of multiple terms) and $\mathbf{I_q} \subseteq \mathbf{I}$ is a set of images that have been clicked for $\mathbf{q}$ in a search engine. Availability of such data implies the existence of (i) text data associated with some or all of the images in $\mathbf{I}$ (e.g., image captions, text in Web pages containing the image, or user-generated tags) and (ii) a search engine where the textual description associated with each image is indexed and searched for by users using textual queries.

3.1 Problem definition

A *concept* c corresponds to a clearly defined, non ambiguous entity and is represented by a set $\{N_c, K_c, D_c\}$, where N_c is the concept's short *name*, K_c are *keywords* that are conceptually related to c, and D_c is a free-text, short *description* of c. An example is the concept with $N_c = traffic$, $K_c = \{$traffic jam, cars, road, highway$\}$ and description $D_c =$ "Image showing a high density of vehicles on a road or highway".

Given an image collection $\mathbf{I}$, our aim is to apply a method m that automatically generates for each concept c a training set $\mathbf{T}_{c,m}$ to be used in a supervised machine learning setting. To this end, method m needs to find a set $\mathbf{I}_{c,m}$ of images that contain the concept c (positive examples), as well a set $\mathbf{I}_{\bar{c},m}$ (disjoint to $\mathbf{I}_{c,m}$) that consists of images that do not contain c (negative examples). This work investigates methods based on clickthrough data collected in search logs to produce the set $\mathbf{I}_{c,m}$. The generation of $\mathbf{I}_{\bar{c},m}$ is based on random selection.

3.2 Search log based positive sample selection

The simplest method for selecting positive samples for a concept c based on search log data is to consider the images that have been clicked for queries that *exactly match* the concept's name N_c; this constitutes method m denoted as *exact*. Clickthrough data though are sparse [7], since (i) images that are relevant may not have been clicked in the past, and (ii) users with the same information need tend to submit different textual queries even when seeking images that are conceptually similar. For instance, images that have been clicked for query $q_1 =$ "building" will be considered by method *exact* as positive samples for $N_c = building$, whereas images clicked for query $q_2 =$ "tall building" will not, since q_2 does not exactly match N_c. Exact match is therefore bound to produce a relatively small number of samples per concept; to address this, methods with less stringent criteria for matching queries to concepts are proposed next.

For each image, the terms in the queries for which the image has been clicked are used in order to create a surrogate textual description for that image (similar to [24]). This can then be viewed as a document (in the traditional IR sense) that can be indexed and retrieved in response to a query. To this end, we employ a *language modelling* (LM) approach to IR [12]. In this approach, a language model φ_D is inferred for each document D. Given query Q, the documents are ranked by estimating the *likelihood of the query* $P(Q|\varphi_D)$. Queries are represented as sequences of k binary random variables each corresponding to a term, and the query likelihood is:

$$P(\mathbf{q}|\varphi_D) = P(q_1, q_2, \ldots, q_k|\varphi_D) = \prod_{i=1}^{k} P(q_i|\varphi_D) \tag{1}$$

assuming that each q_i is generated independently from the previous ones given the document model. The language model is thus reduced to modelling the distribution of each single term. The simplest estimation strategy for an individual term probability is the *maximum likelihood estimate* (*mle*). This corresponds to the relative frequency of a term t_i in document d, $P_{\text{mle}}(t_i|\varphi_d) = \frac{tf_{i,d}}{\sum_t tf_{t,d}}$, where $tf_{i,d}$, the term frequency of term t_i in d, is normalised by the document's length (the sum of the term frequencies of all of its terms). This method for selecting positive samples for concept c is denoted as LM when we use the concept name N_c as the query, and as LM_{key} when we use the concept name N_c together with the concepts' keywords K_c as the query. Using this method, images I that have, for instance, been clicked for query "tall building", but not for query "building", will be selected as positive samples for $N_c =$ *building* since their $P(N_c|\varphi_I) > 0$.

Equation (1) assigns zero query likelihood probabilities to documents missing even a single query term. This sparse estimation problem is addressed by *smoothing* techniques, that redistribute some of the probability of the terms occurring in a document to the absent ones. We use a mixture model of the document model with a background model (the collection model in this case), well-known in text retrieval as Jelinek–Mercer smoothing [12]:

$$P(\mathbf{q}|\varphi_D) = \prod_{i=1}^{k} (1-\lambda) P_{\text{mle}}(q_i|\varphi_D) + \lambda P_{\text{mle}}(q_i|\varphi_C) \tag{2}$$

where λ is a smoothing parameter (typically set to 0.8), and $P_{\text{mle}}(t_i|\varphi_C) = \frac{df_i}{\sum_t df_t}$, with df_i the document frequency of the term t_i in the collection. In our case the collection consists of the images that appear in the clickthrough data, i.e., images that have been previously clicked for some query. The selection method based on this smoothed LM is denoted as LMS when the concept name N_c is used as the query, and as LMS_{key} when the concept name N_c together with the concepts' keywords K_c are used as the query. For $N_c =$ *rally motorsport*, for example, this method will allow for the selection as positive samples of images I that have been clicked for queries containing the term "rally", but have never been clicked for queries that contain the term "motorsport".

The aim of these four LM-based selection strategies is to increase the number of positive samples by progressively relaxing the strictness of the matching criteria. This can be further achieved by applying stemming in each of these methods, resulting in LM_{stem}, $LM_{\text{key\_stem}}$, LMS_{stem}, and $LMS_{\text{key\_stem}}$, respectively. The open source `PF/Tijah`[6] retrieval system [13] is used as the implementation of the above retrieval approaches.

The final technique exploits the clickgraph in order to deal with the data sparsity and the possible mismatch of users' query terms to concept names and keywords. The basic premise of this approach is that images clicked for the same query are likely to be relevant to each other, in the sense that their visual content is likely to pertain to similar semantic concept(s). For each concept c, an initial image set that contains the images selected using the *exact* method is constructed. If this method does not produce any results, we add the images using the LM retrieval model (i.e., the images clicked for the most textually similar query to the concept name). This initial image set is then expanded with the images accessible by a 2-step traversal of the graph as follows. First, each image i in this initial set is added to a final set. For each such i, the queries for which this image was clicked are found, and, then, for each such query, the images clicked for that query are added to the final set (other than the ones already there). Consider, for instance, image I_1 that has been clicked both for query "fire" and for query "flames", and image I_2 that has been clicked only for query "flames". For concept $N_c = \mathit{fire}$, both these images will be added in the final set. This method is denoted as *clickgraph* and produces a set of images. To rank these images, one approach is to apply this method after assigning weights to the edges of the clickgraph based on the number of clicks. Alternative approaches that exploit the clickgraph are iterative methods, such as the random walk models employed in [7].

Even though methods such as the ones described above aim to deal more effectively with the sparsity of the clickthrough data, they are likely to introduce false positives in the sample selection, i.e., images that were clicked but were not relevant. Given that our proposed methods produce a ranking of the images, a strategy to reduce this potential noise would be to filter the selected images by considering only those ranking above a given threshold. In this work, however, such noise reduction techniques are not applied; all retrieved images are considered as positive samples (for the LM-based methods, all samples with $P(Q|\varphi_D) > 0$ are selected as positive).

3.3 Negative sample selection

Negative samples are selected randomly. The probability of selecting a non-negative (i.e., positive) example in the original dataset $\mathbf{I}$ is equal to the concept's prior probability in $\mathbf{I}$, i.e., $P(c|\mathbf{I})$. Assuming that after positive sample selection the prior of c in the remaining set decreases, $P(c|\mathbf{I} - \mathbf{I}_{c,m}) \leq P(c|\mathbf{I})$, then the prior $P(c|\mathbf{I})$ is an upper bound for the probability of error. Random negative sample selection will therefore be accurate for rare concepts.

The number of negative examples has to be sufficient for training (e.g., description of the class boundaries in minimum margin classifiers). At the same time, though, it should not be too high, since that would lead to an increase in the number of false negatives. In this work, the number of negatives are abritrarily set to $N_{\bar{c},m} =$

[6] http://dbappl.cs.utwente.nl/pftijah/

$\max(1{,}000 - N_{c,m}, N_{c,m})$, where $N_{c,m} = |\mathbf{I}_{c,m}|$ is the number of positive examples for c. This approach ensures that, for any concept, its training set contains at least 1000 samples in total and that its prior in this set is below 0.5. In case the number of positive examples is high (above 500), then the number of negative examples increases accordingly, so that enough samples are available for the possibly more complex classification/ranking problem that arises.

3.4 Automatic image annotation

For each image in the collection, two types of low-level features are extracted, one capturing visual information in the image and another based on text captions accompanying the images. Both features are similar to the ones used in [29]. Text features are required since some concepts cannot be described using visual features only (e.g., "war"). Using features based on text allows the evaluation of the generated training sets for these concepts. In any case, however, relevance judgments are based on the visual content of images and not on their metadata.

For the visual description, the Integrated Weibull distribution [10] is extracted from a number of overlapping image regions. The region distributions are then compared against the distributions of images belonging to a set of common reference concepts (or proto-concepts). This leads to a $120 - d$ feature vector $\mathbf{F}_W$.

For the text-based feature vector, a vocabulary of the most frequently used words is built for each concept, using the available textual metadata. Each image caption is compared against each concept vocabulary and a frequency-histogram $\mathbf{F}_{T,c}$ is built for each concept c. The feature vector length is equal to the vocabulary size, but is usually very sparse due to the short length of typical captions.

For ranking with classifiers, each image is represented by its (visual or text-based) feature vector and the score output by a support vector machine (SVM) classifier. The classifiers employ an RBF kernel and 3-fold cross-validation is performed on the training set to select the class weight parameters $w+$ and $w-$. The LibSVM [4] implementation is used as the basis of the classification system.

These classifiers and low-level features are adopted due to the effectiveness they have demonstrated for the task of automatic annotation of multimedia content [29, 39]. Even though more sophisticated or more heavily tuned methods are probably more effective, their application is beyond the scope of this paper. Our aim is not to optimise the effectiveness, but to compare concept classifiers trained on manually labelled samples to classifiers that consider search log based annotated samples as training data.

4 Experimental design

4.1 Datasets

The image collection $\mathbf{I}$ we use consists of 97,628 photos provided by *Belga News Agency*[7] in the context of the activities of the VITALAS [8] project. The photographic

[7] http://www.belga.be/

[8] http://vitalas.ercim.org/

images cover a broad domain, and can be characterised either as “editorial”, i.e., pictures with concrete content related to a particular event, e.g., sports, politics, etc., or as “creative”, i.e., pictures with artistic and timeless content, such as nature, work, etc. Each photo is accompanied by high quality metadata (defined by the IPTC[9]) that include textual captions written manually by Belga’s professional archivists.

Belga also provided us with their search logs for a period of 101 days from June to October 2007. From these, we extracted the clickthrough data and performed a “light” normalisation on the text of the submitted queries, so as to clean up the data and identify identical/similar queries that had been submitted with slight variations. This preprocessing step included conversion to lower case and removal of punctuation, quotes, the term “and”, and the names of the major photo agencies that provide their content to Belga (e.g., EPA). The normalisation was deliberately kept shallow so that further steps, such as stemming and stopword removal, can be applied at a later stage where required. These search log data contain 35,894 of the images that also belong to **I** and which have been clicked for 9,605 unique queries. Given that Belga is a commercial portal, their search log data are much smaller in size, compared to those collected, for instance, by a general purpose search engine [7]. On the other hand, given that it provides services to professional users, mainly journalists, we expect their search log data to be relatively less noisy. The sparsity of the clickgrough data is evident, though, similarly to [7], in the power-law distributions observed for the images-per-query and queries-per-image pairs.

The VITALAS project has developed a multimedia concept lexicon which currently stands at around 600 entries. These concepts have been selected following a multi-step process involving a statistical analysis of Belga’s image captions [23], feedback by Belga’s professional archivists, and the addition of concepts from the MediaMill [29] and LSCOM [20] lexicons. Out of these, we selected 25 concepts for our experiments (see Table 1) based on various criteria, including the availability of search log based positive samples and whether they are generalisable across collections. We also aimed to include a large number of sports-related concepts, given that 38.8% of the images in **I** have been classified as belonging to the IPTC subject “sport”. Given the manual annotations described next, we also aimed to include concepts with high variation in their frequencies in the manually annotated sets.

A large-scale manual annotation effort has been undertaken by Belga staff for the images in collection **I**. The presence of the VITALAS concepts was assumed to be binary. This process has yielded an incomplete, but reliable ground truth. For our selected 25 semantic concepts c, their manual annotation sets contain between 994 and 1,000 annotated samples.

Table 2 lists the number of positive samples for each of the methods employed to generate a training set. Missing values indicate that no positive samples could be acquired for that concept-method combination. The fact that method *exact* not only generates small numbers of positive samples, but also does so for a small number of concepts (9 out of 25), illustrates the need for the application of methods with less strict matching criteria. As expected, the number of positive samples increases with

[9] http://www.iptc.org/

Table 1 The list of the 25 VITALAS concepts used in the experiments together with their keywords

Concept c		
ID	Name	Keywords
1	airplane_flying	air
2	airport	plane, runway
3	anderlecht	sport, soccer, football, club, belgian
4	athlete	sport
5	basketball	nba, competition, team, dribbling, passing, sport, player
6	building	
7	club_brugge	soccer, football, game, match, breydel, player, club bruges, dexia, belgian
8	crowd	mass, event, protest, demostration, people
9	farms	agricultural, people, field, countryside
10	fashion_model	
11	fire	red flames, warm, fireman, firefighter
12	flood	rain, river
13	formula_one	f1, ecclestone, ferrari, mclaren, bmw, raikkonen, hamilton
14	highway	road, freeway, superhighway, autoroute, autobahn, expressway, motorway
15	logo	
16	meadow	sheep, goats, grass, field
17	rally_motorsport	motor, racing
18	red_devils	sport, soccer, football, belgian
19	sky	clouds, sun, moon
20	soccer	football
21	stadium	sport, game, match, competition, athleticism, stands, tracks
22	team	group
23	tennis	racket, court, match
24	volleyball	volley, ball, net, beach
25	war	

stemming, smoothing, and addition of keywords. For the *clickgraph* method, there is no apparent trend when compared to the other methods; the number of positive samples appears to be very concept-specific. Overall, the manually annotated data contain on average more positive samples than the search log based ones.

4.2 Description of experiments

Four types of experiments are performed for evaluating the effectiveness of using search log based methods to select images to be used as alternative or complementary data sources to manual annotations for training concept classifiers. Our experimental setting refers to the search logs as 'SL', the manually annotated set as 'MA', and the common evaluation set (defined in Experiment 2) as 'CE'. Below is a detailed description the experiments while Table 3 provides a summary of their most important attributes.

4.2.1 Experiment 1: SL training, MA evaluation (feasibility test)

This experiment is a first indication on the usefulness of the automatically generated training sets $\mathbf{T}_{c,m}$. For this experiment, the classifiers are built using the training data originating from the search logs $\mathbf{T}_{c,m}$ (as described in Section 3) that do not overlap

Table 2 Number of positive samples $\mathbf{N}_{c,m}$ per training set generation method m for each concept c

c	$\mathbf{N}_{c,m}$								
	manual	*exact*	*LM*	LM_{stem}	*LMS*	LMS_{stem}	LMS_{key}	$LMS_{\text{stem\_key}}$	*clickgraph*
1	26						13	13	
2	217		66	66	66	66	69	69	47
3	322	212	282	282	282	282	1,452	1,458	4,927
4	396			23		23	6	35	
5	384	7	15	15	15	15	63	70	7
6	103		26	26	26	26	26	26	116
7	408	31	53	53	165	165	1,272	1,287	3,610
8	149						42	42	
9	9						51	59	
10	432				70	71	70	71	
11	97	5	65	65	65	65	271	285	169
12	362			4		4	19	20	969
13	420	9	20	56	58	56	164	162	84
14	162						26	27	
15	430								55
16	1						21	21	
17	276		9	9	9	9	12	26	16
18	461	23	41	42	114	114	1,196	1,202	3,039
19	145						23	27	
20	428	329	935	935	935	935	964	964	6,233
21	109		12	12	12	12	28	49	
22	37		28	28	28	28	32	28	102
23	371	6	24	24	24	24	41	41	533
24	340	23	43	43	43	43	87	87	710
25	207		7	7	7	7	7	7	18
Mean	251.68	71.67	108.4	99.41	119.94	108.06	248.13	253.17	1,289.69
Median	276	23	28	28	50.5	35.5	41.5	45.5	142.5
σ	156.57	116.82	238.35	224.58	228.77	217.67	454.01	455.07	2,009.53

with the manual annotations. Image representation is based on the $\mathbf{F}_W$ features only, i.e., only visual information is used. For results to be comparable across the different positive sample selection methods m, the negative sample set is the same across all datasets of each concept. Effectiveness is measured on the data already manually annotated by Belga's archivists. This allows us to directly compute the evaluation metrics, without performing any manual assessments, but results in each concept having its own evaluation set.

Table 3 Summary of the experimental setup

Experiment	Training set		Evaluation set	Description
1	$\mathbf{T}_{c,m} - \mathbf{T}_{c,\text{manual}}$	(SL − MA)	$\mathbf{T}_{c,\text{manual}}$	Feasibility test
2	$\mathbf{T}_{c,m}$	(SL)	$\mathbf{I}_{\text{eval}}$	Effectiveness of SL alone
3	$\mathbf{T}_{c,m} \cup \mathbf{T}_{c,\text{manual}}$	(SL + MA)	$\mathbf{I}_{\text{eval}}$	Effectiveness of both SL and MA
4	$\mathbf{T}_{c,\text{manual}}$	(MA)	$\mathbf{I}_{\text{eval}}$	Effectiveness of MA (baseline)

'SL' refers to search logs, 'MA' to manual annotations, and $\mathbf{I}_{\text{eval}} = \mathbf{I} - \bigcup_{i,j}(\mathbf{T}_{c_i,m_j} \cup \mathbf{T}_{c_i,\text{manual}})$

4.2.2 Experiment 2: SL training, CE evaluation

This experiment provides an evaluation on the effectiveness of using the automatically generated training data alone. The evaluation set $\mathbf{I}_{\text{eval}}$ for this experiment is common for all concepts (and is also the same for the following Experiments 3 and 4). It is obtained after removing all manual and automatically generated sets from the original image set $\mathbf{I}$. Hence, it is the set $\mathbf{I}_{\text{eval}} = \mathbf{I} - \bigcup_{i,j}(\mathbf{T}_{c_i,m_j} \cup \mathbf{T}_{c_i,\text{manual}})$ which contains 56,605 images. Note that for a given concept c, the randomly selected negative examples are common for all methods m. In this experiment the training sets are generated using the search logs (i.e., no manual annotations are used) and evaluation is performed on the common evaluation set. Visual or text-based low level features are used.

4.2.3 Experiment 3: SL & MA training, CE evaluation

This experiment examines the effect of combining the training sets of Experiment 2 with the manually annotated data. Hence new training sets are generated for each non-manual method m, such that $\mathbf{T}'_{c,m} = \mathbf{T}_{c,m} \cup \mathbf{T}_{c,\text{manual}}$. If an image belongs to both $\mathbf{T}_{c,m}$ and $\mathbf{T}_{c,\text{manual}}$ the manual annotation takes priority. Note that both the randomly selected and the manually annotated negative examples are used. Evaluation is again performed on the common evaluation set $\mathbf{I}_{\text{eval}}$ and classification uses either visual or text-based low-level features.

4.2.4 Experiment 4: MA training, CE evaluation

This is a baseline experiment, where only manual annotations are used to train the classifiers, which are then evaluated on $\mathbf{I}_{\text{eval}}$ for visual or text-based features. Generally, manual annotations are expected to provide the best results since they contain the most accurate and reliable assessments resulting in training sets of higher quality.

5 Experimental results

5.1 Effectiveness

Results for the first experiment (feasibility test) are directly produced by using the existing manual annotations as ground truth for the evaluation. Table 4 shows the results in terms of the average precision (AP) attained for each concept and training set generation method (this is averaged over ten runs so as to avoid bias due to random negative sample selection) and the mean average precision (MAP) achieved over all concepts for each of these methods.

The AP values though are not comparable across concepts, given that the lower bound of the AP for a concept, i.e., the AP of a "random classifier", is not zero, but corresponds to the prior of that concept in the test set [39]. Therefore, the classifier for each concept should be evaluated on how much it improves over the prior by using, for instance, $\Delta\text{AP} = \text{AP} - \text{prior}$, i.e., the difference in AP between the concept's classifier and a random classifier in the same dataset. To make this

Table 4 Average precision (AP) for each concept *c* and training set generation method *m*, mean AP and mean ΔAP across concepts, together with the prior for each concept for Experiment 1 (feasibility test)

c	$\mathbf{T}_{c,m}$								Prior
	exact	*LM*	LM_{stem}	*LMS*	LMS_{stem}	LMS_{key}	$LMS_{\text{stem\_key}}$	*clickgraph*	
1						0.06	0.06		0.03
2	0.23	0.37	0.37	0.37	0.37	0.36	0.36	0.30	0.22
3	0.59	0.60	0.60	0.60	0.60	0.55	0.55	0.52	0.32
4			0.45		0.45	*0.34*	0.49		0.40
5	0.52	0.55	0.55	0.55	0.55	0.55	0.55	0.52	0.39
6		0.22	0.22	0.22	0.22	0.22	0.22	*0.08*	0.10
7	0.54	0.58	0.58	0.52	0.52	0.61	0.61	0.50	0.41
8						0.39	0.39		0.15
9						0.06	0.07		0.01
10				0.72	0.71	0.72	0.71		0.43
11	0.39	0.45	0.44	0.45	0.44	0.34	0.34	0.36	0.10
12	0.52	0.58	0.59	0.58	0.59	0.53	0.45	0.43	0.36
13	0.43	0.55	0.58	0.59	0.58	0.72	0.73	0.73	0.42
14						0.31	0.32		0.16
15								0.55	0.43
16						0.02	0.02		0.00
17		0.69	0.69	0.69	0.69	0.72	0.70	0.56	0.28
18	0.75	0.78	0.80	0.80	0.80	0.72	0.72	0.66	0.46
19						0.27	0.24		0.15
20	0.57	0.64	0.64	0.64	0.64	0.65	0.65	0.56	0.43
21		0.40	0.40	0.40	0.40	0.22	0.19		0.11
22		0.22	0.22	0.22	0.22	0.19	0.22	0.07	0.04
23	0.45	0.50	0.50	0.50	0.50	0.50	0.50	0.46	0.37
24	0.56	0.57	0.57	0.57	0.57	0.51	0.51	0.45	0.34
25		*0.17*	*0.17*	*0.17*	*0.17*	*0.17*	*0.17*	0.31	0.21
Mean AP	0.50	0.49	0.49	0.51	0.50	0.41	0.41	0.44	0.25
Mean ΔAP	0.16	0.21	0.20	0.21	0.20	0.16	0.16	0.14	0.00

The cases where the classifier performs worse than random, i.e., ΔAP is negative, are highlighted in *italics*

comparison possible, Table 4 also lists the prior of each concept in the test set and the mean ΔAP over all concepts for each training set generation method.

The results for Experiment 1 indicate that, in most cases, the AP value is considerably higher than the prior. The only cases where the classifier performs worse than random, i.e., ΔAP is negative, are: (i) the LMS_{key} training set generation method for concept *athlete* (#4), for which either the addition of the much broader keyword "sport" leads to topic drift and, therefore, to the inclusion of false positives, or the classifier is not able to produce effective results given the small number (6) of positive samples for this concept-method combination, (ii) the clickgraph method for concept *building* (#6), where the many more positive samples generated by this method compared to the LM-based ones appear to contain a lot of noise, and (iii) all LM-based methods for concept *war* (#25) which produce small numbers (7) of positive samples for a concept already considered to be "difficult", since it can be argued that it is too high level so as to be detected from visual information alone. The results for Experiment 1 further indicate that training set generation methods based on language modelling tend to perform better than the exact match approach, leading to the conclusion that the additional samples obtained are useful, whereas the clickgraph method performs worse than the LM approaches, despite the increased number of samples; this can be attributed to the noise that this method introduces.

For Experiments 2, 3 and 4, ground truth is not available for the common evaluation set. In order to assess the ranking performance, the authors manually annotated for each concept the set created by pooling the top 200 results of all experiments for that concept. As evaluation metric, the precision at the first 200 results (P@200) is used.

Figure 1 provides the mean P@200 across all concepts for all features and training set generation methods, and allows the following interesting observations. (i) For $\mathbf{F}_W$, the automatically generated training data alone (exp. 2) cannot surpass the performance of the manually produced ones (exp. 4). (ii) Combining the two training data sources, however, consistently gives the best results for $\mathbf{F}_W$ (exp. 3). (iii) Surprisingly, the use of $\mathbf{F}_{T,c}$ in Experiment 2 results in the less noisy methods (the ones not involving keywords or the clickgraph) producing better results compared to methods based on the inclusion of manual annotations (exps. 3 and 4). (iv) Regarding the comparison between the low-level features, $\mathbf{F}_{T,c}$ dominates, but this is to be expected. Examination of the results per concept, however, reveals that in some cases (e.g., for concepts *sky* (#19) and *crowd* (#8)), $\mathbf{F}_W$ achieves better performance. This is typically observed for concepts strongly associated with the image content, rather than the image context. This is also illustrated in Fig. 2 that shows a detailed per concept example for one training set generation method.

Table 5 presents the maximum P@200 achieved for each concept with the corresponding training set generation method. This table and especially the results of Experiment 2 provide a confirmation of our previous indication that the language modelling methods produce better results than the exact match and clickgraph approaches. In addition, Experiment 3 generally improves the results over ones obtained from the manual annotations. The contribution from the search log based training data is therefore positive.

Figure 3 provides a more qualitative view of the results by illustrating samples taken from the manual and automatically generated training sets, as well as the

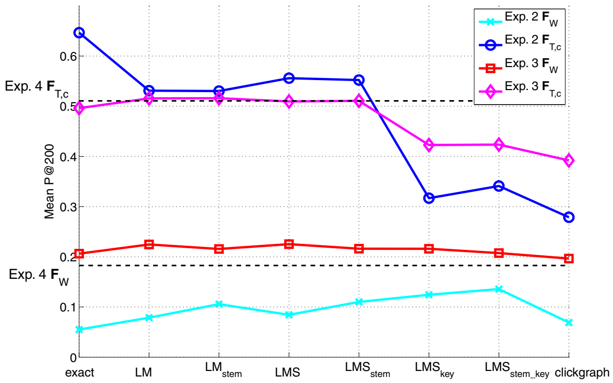
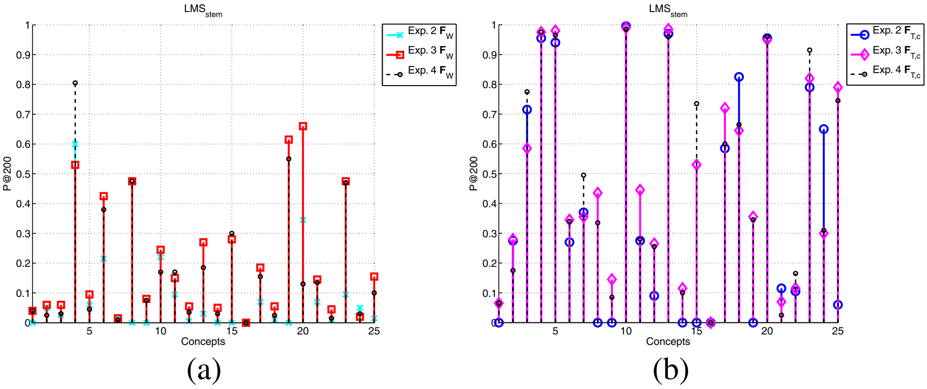


Fig. 1 Mean of P@200 across all concepts for all experiments and training set generation methods



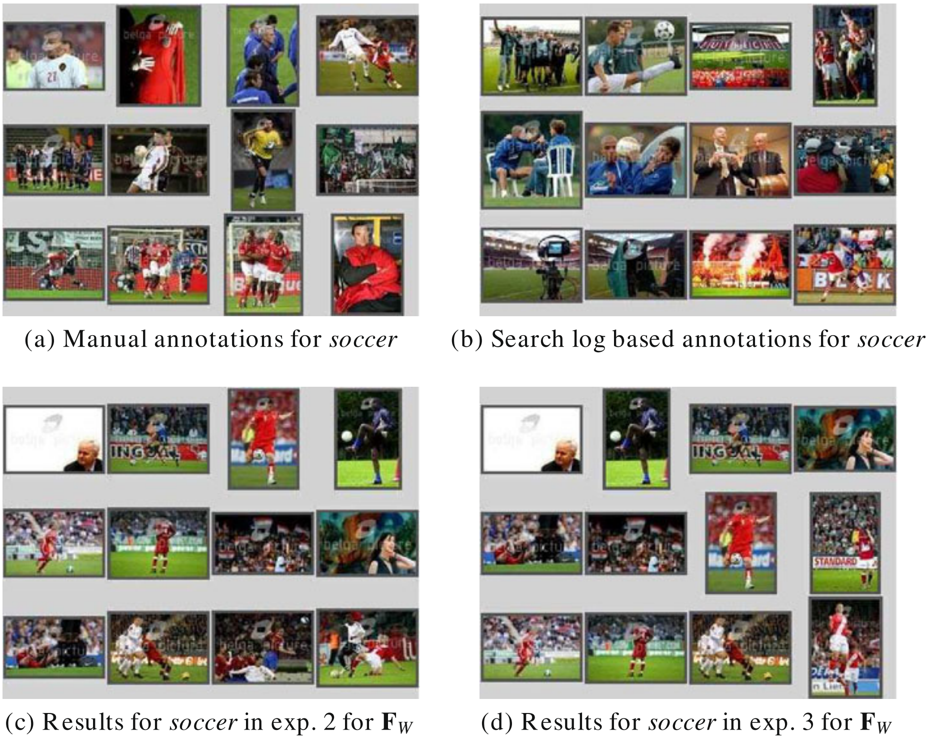
(a) (b)

Fig. 2 An example of the $P@200$ values attained per concept for $\mathbf{T}_{LMS_{\text{stem}}}$ in all experiments and low-level features. A value of $P@200 = 0$ for both features indicates that the corresponding concept has not been evaluated due to insufficient training data for method LMS_{stem}

Table 5 Maximum P@200 value achieved and the corresponding training set generation method per concept c

c	Experiment 2				Experiment 3				Experiment 4	
	$\mathbf{F}_{W,c}$		$\mathbf{F}_{T,c}$		$\mathbf{F}_{W,c}$		$\mathbf{F}_{T,c}$		$\mathbf{F}_{W,c}$	$\mathbf{F}_{T,c}$
1	0.02	LMS$_{\text{key}}$	*0.06*	LMS$_{\text{key}}$	*0.04*	exact	*0.07*	LMS$_{\text{key}}$	0.04	0.06
2	*0.05*	LM	*0.28*	LMS$_{\text{key}}$	*0.06*	LMS$_{\text{key}}$	*0.28*	LM	0.02	0.17
3	*0.03*	LMS$_{\text{key}}$	0.71	LM	*0.06*	exact	0.60	exact	0.03	0.77
4	0.60	LM$_{\text{stem}}$	0.95	LM$_{\text{stem}}$	0.76	LMS$_{\text{key}}$	*0.97*	exact	0.80	0.97
5	*0.06*	LM	0.94	LM	*0.10*	LMS$_{\text{stem\_key}}$	*0.98*	exact	0.04	0.96
6	0.21	LM	0.27	LM	*0.42*	LM	*0.34*	exact	0.38	0.34
7	*0.03*	LM	0.44	LM	*0.01*	LM	0.49	LM	0.01	0.49
8	*0.49*	LMS$_{\text{key}}$	0.32	LMS$_{\text{key}}$	*0.47*	exact	*0.45*	LMS$_{\text{key}}$	0.47	0.33
9	*0.12*	LMS$_{\text{stem\_key}}$	*0.13*	LMS$_{\text{stem\_key}}$	*0.08*	exact	*0.14*	exact	0.07	0.08
10	*0.22*	LMS$_{\text{stem}}$	*1.00*	LMS	*0.25*	exact	*0.99*	LMS	0.17	0.98
11	0.11	exact	*0.31*	exact	*0.17*	exact	*0.44*	LM	0.17	0.28
12	0.01	LMS$_{\text{key}}$	0.15	LMS$_{\text{key}}$	*0.06*	LMS$_{\text{key}}$	*0.26*	exact	0.03	0.25
13	0.16	LMS$_{\text{key}}$	*0.97*	LM	*0.28*	exact	*0.98*	LM$_{\text{stem}}$	0.18	0.96
14	0.02	LMS$_{\text{key}}$	*0.11*	LMS$_{\text{stem\_key}}$	*0.05*	exact	*0.11*	exact	0.03	0.10
15	*0.37*	clickgraph	0.29	clickgraph	0.28	exact	0.58	clickgraph	0.30	0.73
16	*0.09*	LMS$_{\text{key}}$	*0.02*	LMS$_{\text{key}}$	*0.08*	LMS$_{\text{key}}$	*0.02*	LMS$_{\text{key}}$	0.00	0.00
17	0.08	clickgraph	*0.67*	LMS$_{\text{key}}$	*0.19*	LMS$_{\text{stem\_key}}$	*0.72*	LM	0.15	0.60
18	*0.03*	LM$_{\text{stem}}$	*0.82*	LM$_{\text{stem}}$	*0.05*	LMS	0.64	LMS	0.02	0.66
19	0.39	LMS$_{\text{stem\_key}}$	0.10	LMS$_{\text{key}}$	*0.61*	exact	*0.35*	exact	0.55	0.34
20	*0.54*	LMS$_{\text{key}}$	*0.98*	LMS$_{\text{key}}$	*0.66*	LMS$_{\text{key}}$	*0.96*	exact	0.13	0.96
21	0.07	LM	*0.11*	LM	*0.15*	exact	*0.08*	exact	0.13	0.02
22	*0.01*	LMS$_{\text{key}}$	0.12	LMS$_{\text{key}}$	*0.04*	exact	0.14	exact	0.01	0.16
23	0.11	clickgraph	0.84	LMS$_{\text{key}}$	*0.49*	LMS$_{\text{key}}$	0.83	exact	0.47	0.91
24	*0.07*	exact	*0.65*	LM	*0.03*	LMS$_{\text{key}}$	*0.31*	LMS$_{\text{key}}$	0.03	0.31
25	*0.01*	LM	0.06	LM	*0.15*	LM	*0.82*	exact	0.10	0.74

The results that reached or exceeded the baseline results of the manual annotations are highlighted in *italics*



(a) Manual annotations for *soccer*

(b) Search log based annotations for *soccer*

(c) Results for *soccer* in exp. 2 for $\mathbf{F}_W$

(d) Results for *soccer* in exp. 3 for $\mathbf{F}_W$

Fig. 3 *Top row* Positive examples from the manual annotations and an automatically generated training set using the $LMS_{\text{stem\_key}}$ method. *Bottom row* The top 12 results for $\mathbf{F}_W$ in Experiments 2 and 3. All images ©Belga (please refer to http://olympus.ee.auth.gr/~diou/civr2009/belga.html for the full copyright notice)

results for a run that uses $\mathbf{F}_W$. Readers are invited to visit http://olympus.ee.auth.gr/∼diou/civr2009/ to view the image lists returned for each concept and method combination, along with the training set used and the performance achieved.

Overall, the experimental results indicate that the contribution of search log based samples for training concept classifiers is positive, particularly when combined with manually annotated samples. There are also several cases where the classifiers trained on the automatically generated data outperform those built on manually labelled samples, e.g., the classifiers built for concepts *airport* (#2), *farms* (#9), *fashion_model* (#10), *meadow* (#16), *red_devils* (#18), *soccer* (#20), and *volleyball* (#24), irrespective of whether visual or text-based features are used (see Table 5). These are interesting observations given that automatically acquired search log based labelled data are inherently noisy compared to the reliable manual annotations. To provide insights into the interplay between the noise in the training data and the effectiveness of classifiers built using these training data, the remainder of this section first examines the reliability of the search log based training samples employed in our experiments (Section 5.2) and then the effect of the reliability of the training data on the effectiveness of the classifier (Section 5.3).

5.2 Reliability of search log based training samples

To determine the extent to which the assumption forming the basis for the use of search log based annotated samples for training concept classifiers holds, i.e., the assumption that queries for which an image was clicked can be considered as annotations of the image's visual content, this section examines the reliability of the positive samples generated by the search log based methods as image annotations by comparing them against manual annotations; these manual annotations for the search log based sets were performed by the authors. For each concept and training set generation method, the reliability of the sample generated by a method for a concept is defined as the *sample precision*, i.e., the number of true positives in the sample when compared to the manual annotations for that concept.

Table 6 presents for each concept the sample precision of the training sets generated by each of the search log based methods. As expected, the most reliable samples are produced by method *exact*. The sets generated by the LM-based methods that do not take into account the keywords are also very reliable, given that they include additional samples compared to the sets generated by method *exact*, with

Table 6 Sample precision per training set generation method m for each concept c

c	$\mathbf{I}_{c,m}$							
	exact	*LM*	LM_{stem}	*LMS*	LMS_{stem}	LMS_{key}	$LMS_{\text{stem\_key}}$	*clickgraph*
1						0.08	0.08	
2		0.70	0.70	0.70	0.70	0.67	0.67	0.47
3	0.65	0.65	0.65	0.65	0.65	0.15	0.15	0.06
4			1.00		1.00	0.17	0.86	
5	0.71	0.87	0.87	0.87	0.87	0.32	0.29	0.71
6		0.85	0.85	0.85	0.85	0.85	0.85	0.14
7	0.68	0.79	0.79	0.52	0.52	0.07	0.07	0.03
8						0.36	0.36	
9						0.37	0.44	
10				0.90	0.90	0.90	0.90	
11	1.00	0.80	0.80	0.80	0.80	0.19	0.18	0.44
12			0.75		0.75	0.32	0.35	0.00
13	0.89	0.65	0.61	0.59	0.61	0.80	0.81	0.90
14						0.46	0.44	
15								0.53
16						0.43	0.43	
17		1.00	1.00	1.00	1.00	1.00	0.50	0.25
18	0.87	0.93	0.93	0.57	0.57	0.09	0.09	0.04
19						0.22	0.19	
20	0.78	0.81	0.81	0.81	0.81	0.80	0.80	0.42
21		0.67	0.67	0.67	0.67	0.29	0.16	
22		0.68	0.68	0.68	0.68	0.59	0.68	0.05
23	1.00	0.92	0.92	0.92	0.92	0.56	0.56	0.31
24	0.96	0.98	0.98	0.98	0.98	0.61	0.61	0.07
25		0.29	0.29	0.29	0.29	0.29	0.29	0.11
Mean	0.84	0.77	0.78	0.74	0.75	0.44	0.45	0.28
Median	0.87	0.80	0.80	0.75	0.78	0.37	0.44	0.20
σ	0.14	0.18	0.18	0.19	0.19	0.28	0.27	0.27

an average sample precision of over 0.75. The reliability of the samples generated by the keyword-based LM methods and by method *clickgraph* is considerably lower on average, although there exist concepts for which these methods produce reliable samples, e.g., concept *rally_motorsport* (#17) for method LMS_{key}, albeit for only 12 samples (see Table 2), and concept *formula_one* (#13) for method *clickgraph*.

Figure 4 shows for each method the total number of concepts for which that method generated samples and the distribution of those concepts across various levels of sample precision. Overall, the sample precision varies significantly across concepts, particularly for the keyword-based LM methods and method *clickgraph*. For the remaining methods, the sample precision for the majority of the concepts is over 0.6, with half of the concepts having samples with precision over 0.8.

Other studies [9, 16, 25, 28] have also previously examined the reliability of clickthrough data as implicit user feedback. For clickthrough data obtained from text search engines, one study found that 39% of the clicked documents were deemed to be relevant and 29% partially relevant [9], while another concluded that only 52% of the clicked documents are indeed relevant [25]. This low reliability coupled with the observation that users' clicking behaviour is influenced by the overall quality of retrieval results and the order in which these results are presented have cast doubts on the usefulness of clickthrough data as absolute relevance assessments and have instead advocated their usefulness as relative judgments [16]. For the case of image search, clickthrough data are expected to be much more reliable given that they are generated by users interacting with retrieval results consisting of thumbnail images, rather than summary information in the form of text snippets, as is the case in text search. This allows users to make more accurate assessments on the relevance of a particular result before clicking on it. A recent study has indeed found that, on average, 88% of clicked images generated by an approach equivalent to our *exact* method are relevant [28], a markedly higher reliability compared to text-based clickthrough data; this finding is also in line with our analysis.

The analysis presented in this section raises the question of the effect of the reliability of clickthrough-based samples on the effectiveness of concept classifiers that use them for training; this question is examined next.

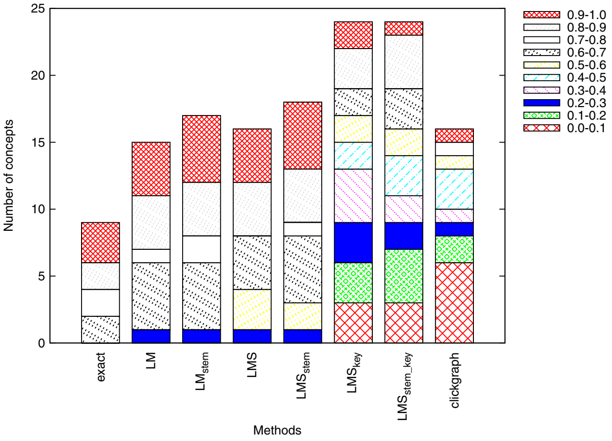
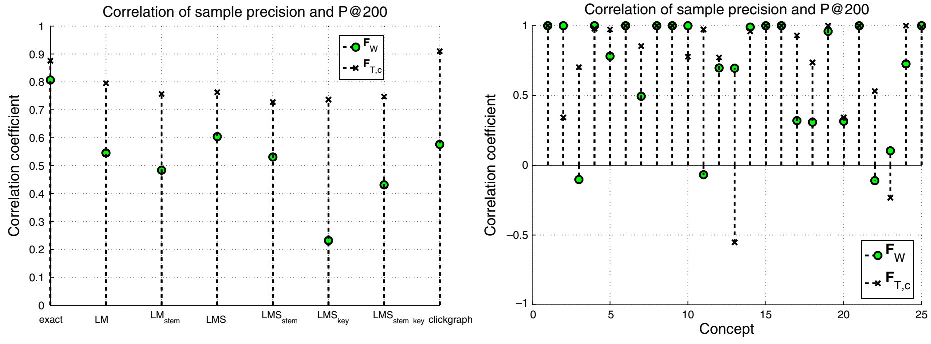


Fig. 4 Frequency distribution of concepts across sample precision levels per training set generation method

5.3 Effectiveness vs. sample precision

To examine the effect of sample precision on the effectiveness, we focus on the results of Experiment 2 which evaluates the P@200 of concept classifiers trained only on search log based samples. Figure 5a shows for each method and low-level feature employed the correlation coefficient between the sample precision and $P@200$. This figure indicates that sample precision and effectiveness are correlated and that, especially for text, this correlation is strong. Given though that each concept's effectiveness depends on many factors, such as the concept's prior in the test set and the difficulty of its detection by the given classifier, providing the correlation coefficient for a given method across concepts cannot lead to safe conclusions. More reliable conclusions can be reached by performing a per concept analysis across all training set generation methods, given in particular that in Experiment 2 the same negative samples are used by all methods for a given concept, and thus the differences in the effectiveness can be directly attributed to the positive samples. Figure 5b shows how the sample precision of the different methods affects the retrieval effectiveness for each concept separately. Again, with few exceptions (that will be discussed in the following) sample precision and effectiveness appear to be strongly correlated. Furthermore, in most cases, the correlation observed for $\mathbf{F}_W$ is lower than that of $\mathbf{F}_{T,c}$; this is due to the lower P@200 values achieved on average by classifiers employing visual descriptors.

For concept *formula_one* (#13), a strongly negative correlation is observed in Fig. 5b. Indeed, observation of the actual sample precision values for that concept shows that while these values display a significant variation ($\sigma = 0.134$), the corresponding variation in the P@200 values is much smaller ($\sigma = 0.018$). In fact, the same maximum effectiveness (P@200 equal to 0.97) is achieved for several different sample precision values: 0.65 for LM, 0.61 for LM_{stem}, 0.81 for $LMS_{\mathrm{stem\_key}}$. Practically, this means that in this case the additional true positive samples do not provide extra information for the classifier (since the classifier already achieves near-perfect retrieval effectiveness). At the same time, methods with high sample precision for that concept (such as the *exact* method, with 0.89) do not achieve higher P@200 compared to methods with lower sample precision, since they do not provide an



(a) (b)

Fig. 5 Correlation coefficient between sample precision and effectiveness (**a**) across concepts for each of the methods and (**b**) across methods for each of the concepts

adequate sample size (9 samples for method *exact* as opposed to 162 samples for method $LMS_{\text{stem\_key}}$).

Similar observations can also be made for the other concepts which display a negative correlation in Fig. 5b, i.e., concepts *anderlecht* (#3), *fire* (#11), *team* (#22), and *tennis* (#23). For these concepts, when a drop in sample precision leads to higher effectiveness values, it is due to the high number of samples that are added. For concept *anderlecht* (#3), for instance, the classifier built from the visual features of the 212 samples generated by method *exact* with sample precision of 0.65 achieves P@200 of 0.025, while method $LMS_{\text{stem\_key}}$ for that concept achieves a higher P@200 of 0.035 despite a lower sample precision of 0.15; this can be attributed to the significantly higher number of samples considered by method $LMS_{\text{stem\_key}}$ (1,458 samples). As a matter of fact, method $LMS_{\text{stem\_key}}$ generates for that concept a higher number of true positives ($219 \approx 1458 \times 0.15$) compared to those generated by method *exact* ($138 \approx 212 \times 0.65$).

The above analysis indicates that more insights on the effect of the training sample noise on the effectiveness can be gained by examining the actual number of true positives generated by each method. Given that computing the correlation between the effectiveness and the number of true positives generated by a method is again only meaningful when performed for each concept separately, we devise the following comparison scheme that can be applied across concepts and methods.

Motivated by the observation that the training set generation methods that have been employed are not completely unrelated to each other, Table 7 lists the pairs of methods m_1 and m_2 that exhibit a subset/superset relation, i.e., the positive samples $\mathbf{I}_{c,m_1}$ generated by method m_1 for concept c are a subset of the positive samples $\mathbf{I}_{c,m_2}$ generated by method m_2. Given the number of samples $\mathbf{N}_{c,m_i}$ generated by each method m_i for a particular concept c and the precision of these samples s_{c,m_i}, we can determine the number of additional samples considered by m_2 compared to m_1 ($\Delta\text{size}_c = \mathbf{N}_{c,m_2} - \mathbf{N}_{c,m_1}$), the number of additional true positives ($\Delta{+}_c = s_{c,m_2} \times \mathbf{N}_{c,m_2} - s_{c,m_1} \times \mathbf{N}_{c,m_1}$) and the number of additional false positives ($\Delta{-}_c = \Delta\text{size}_c - \Delta{+}_c$). By considering m_1 as a baseline, we can then gauge the effect of these additional (true positive and false positive) samples contributed by m_2 by examining the difference observed in the effectiveness between the two methods: $\Delta P@200 = P@200_{m_2} - P@200_{m_1}$. When the majority of samples added by a method are true positives, i.e., $\Delta{+}_c/\Delta\text{size}_c$ is close to 1, it is then expected that the effectiveness will improve. Similarly, if the majority of added samples are false positives, i.e., $\Delta{+}_c/\Delta\text{size}_c$ is close to 0, then it is expected that the effectiveness will deteriorate. These observations form the basis of the comparative analysis performed next, that examines the effect of true positive samples on the effectiveness across concepts and methods.

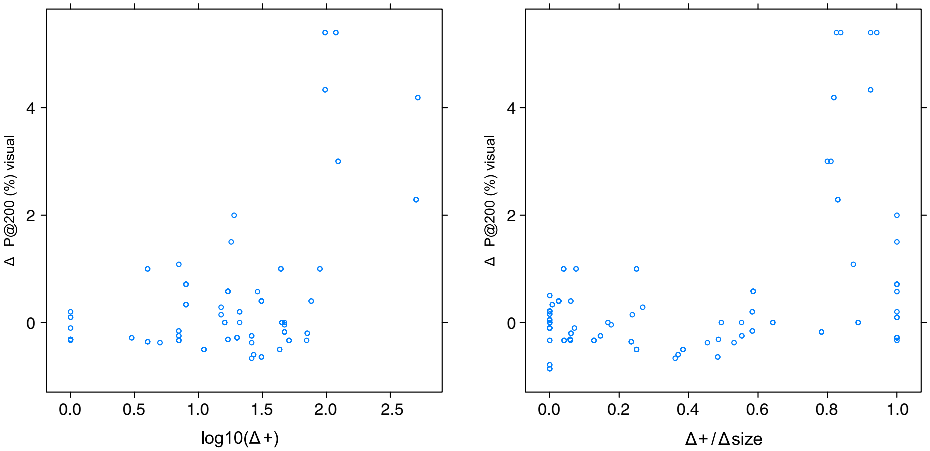
Table 7 Pairs of training set generation methods that exhibit a subset/superset relation

$m_1 \subseteq m_2$ ($\mathbf{I}_{c,m_1} \subseteq \mathbf{I}_{c,m_2}$)		
exact ⊆ LM	LM ⊆ LM_{stem}	$\text{LM}_{\text{stem}} \subseteq LMS_{\text{key\_stem}}$
exact ⊆ LM_{stem}	LM ⊆ LMS	$LMS \subseteq LMS_{\text{stem}}$
exact ⊆ LMS	LM ⊆ LMS_{stem}	$LMS \subseteq LMS_{\text{key}}$
exact ⊆ LMS_{stem}	LM ⊆ LMS_{key}	$LMS \subseteq LMS_{\text{key\_stem}}$
exact ⊆ LMS_{key}	LM ⊆ $LMS_{\text{key\_stem}}$	$LMS_{\text{stem}} \subseteq LMS_{\text{key\_stem}}$
exact ⊆ $LMS_{\text{key\_stem}}$	$\text{LM}_{\text{stem}} \subseteq LMS_{\text{stem}}$	$LMS_{\text{key}} \subseteq LMS_{\text{key\_stem}}$

Given the 25 concepts used in our experiments and the 18 method pairs that exhibit subset/superset relations, we obtain a total of 154 method pairs over all concepts (these are less than all possible combinations since the employed methods do not generate training sets for all concepts). Figure 6a plots the relative difference in the effectiveness $\Delta P@200(\%) = (P@200_{m_2} - P@200_{m_1})/P@200_{m_1}$ vs. the increase in the number of true positives $\Delta+$, while Fig. 6b plots the relative difference in the effectiveness $\Delta P@200(\%)$ vs. the proportion of the added samples that are true positives $\Delta+/\Delta\text{size}$. Figure 6a indicates that on average improvements in the effectiveness are correlated with increase in the number of true positive samples, with significant improvements observed when more than 100 true positive samples are added. Figure 6b indicates the effect of sample precision on the effectiveness by showing that the addition of samples that contain at least 60% true positives tend to result in improvements in the effectiveness.

In both figures, values of $\Delta P@200(\%) < 0$ are observed in some cases. In these cases, the use of new samples has led to a decrease in effectiveness, mostly due to the false positives that have been introduced. In two specific cases, for concepts *flood* and *rally motorsport*, the introduction of 1 and 3 positive examples, respectively, without any false positives, i.e., $\Delta+/\Delta\text{size} = 1$, still led to a decrease in effectiveness (Fig. 6b). This is easily explained, given that both concepts had few positive examples (19 and 9, respectively); with so few positive examples the produced models were not stable and the newly introduced positive examples led the classifier into performing worse.

The results of the analysis presented in this section indicate that (i) there is a strong correlation between sample precision and effectiveness (Fig. 5), (ii) high sample precision alone cannot guarantee that there are benefits from using the samples obtained from clickthrough data (as the examination of individual concept examples indicates), and (iii) there is a high tolerance to noise, as long as there is a large



(a) ΔP@200 (%) F_W vs. Δ+
correlation coefficient = 0.57

(b) ΔP@200 (%) F_W vs. Δ+/Δsize
correlation coefficient = 0.41

Fig. 6 Relative difference in the effectiveness for (**a**) different numbers of additional true positive samples, and (**b**) different proportions of the added samples that are true positives

number of positive samples added (as Fig. 6 indicates). Overall, both sample size and precision affect retrieval effectiveness. Ideally, it would be desirable to obtain a large number of noise-free training samples from the clickthrough data. In practice, however, these are conflicting requirements and a balance between sample precision and size should be preserved.

6 Conclusions & future work

This paper demonstrated how clickthrough data can contribute to reducing the effort required to create and/or maintain the training data needed for automatic image annotation using supervised learning. More specifically, a set of methods have been presented that allow the automatic construction of concept training data given search interaction logs. These methods were tested on a dataset collected from a real-world commercial search engine and the results show that clickthrough data alone can lead to satisfactory effectiveness, while their combination with manual annotations surpasses the effectiveness of using manually generated training data alone.

We expect these results to enable the practical application of the 'detector approach' to annotation, which reduces the investment required to apply image annotation in 'the real world'. Existing content owners can create concept detectors specialised to their domain by simply exploiting the usage logs - or start collecting these right away! The main advantages of our approach grounded in clickthrough data are its scalability in the number of concept detectors, and the possibility to dynamically adapt the detector set, automatically keeping track of concepts that change or emerge.

Regarding the limitations of this approach, the analysis of results identified two factors that are crucial for the effective use of clickthrough data for training set generation: sample size and noise. In the future, we aim at developing tools that will allow the choice of the optimal sample selection method for each concept. Furthermore, we seek ways of reducing the effect of sample noise by developing training strategies that also take into account confidence measures indicating the probability of error of each selected sample. This will allow the application of the proposed methods to diverse and potentially very noisy multimedia search environments, such as the ones encountered on the Web.

Acknowledgements The authors are grateful to the Belga press agency for providing the images and search logs used in this work and to Marco Palomino from the University of Sunderland for the extraction of the text features used. This work was supported by the EU-funded VITALAS project (FP6-045389). Christos Diou is supported by the Greek State Scholarships Foundation (http://www.iky.gr).

References

1. Ashman H, Antunovic M, Donner C, Frith R, Rebelos E, Schmakeit JF, Smith G, Truran M (2009) Are clickthroughs useful for image labelling? In: Pasi G, Bordogna G, Mauri G, Baeza-Yates R (eds) Proceedings of the 2009 IEEE/WIC/ACM international conference on web intelligence (WI 2009), pp 191–197

2. Ayache S, Quénot G (2008) Video corpus annotation using active learning. In: Boughanem M, Berrut C, Mothe J, Soulé-Dupuy C (eds) Proceedings of the 30th European conference on IR research, pp 187–198
3. Baeza-Yates RA, Hurtado CA, Mendoza M (2007) Improving search engines by query clustering. J Am Soc Inf Sci Technol 58(12):1793–1804
4. Chang CC, Lin CJ (2001) Libsvm: a library for support vector machines. Available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm
5. Chang SF, He J, Jiang YG, El Khoury E, Ngo CW, Yanagawa A, Zavesky E (2008) Columbia University/VIREO-CityU/IRIT TRECVID2008 high-level feature extraction and interactive video search. In: Proceedings of TRECVID 2008
6. Chua TS, Tang J, Hong R, Li H, Luo H, Zheng YT (2009) NUS-WIDE: A real-world Web image database from National University of Singapore. In: Marchand-Maillet S, Kompatsiaris Y (eds) Proceedings of the 8th international conference on content-based image and video retrieval (CIVR 2009). ACM Press
7. Craswell N, Szummer M (2007) Random walks on the click graph. In: Proceedings of the 30th ACM SIGIR conference on research and development in information retrieval, pp 239–246
8. Faymonville P, Wang K, Miller J, Belongie SJ (2009) CAPTCHA-based image labeling on the Soylent Grid. In: Bennett PN, Chandrasekar R, Chickering M, Ipeirotis PG, Law E, Mityagin A, Provost FJ, von Ahn L (eds) (2009) Proceedings of the ACM SIGKDD workshop on human computation. ACM Press, pp 46–49
9. Fox S, Karnawat K, Mydland M, Dumais ST, White T (2005) Evaluating implicit measures to improve web search. ACM Trans Inf Syst 23(2):147–168
10. van Gemert JC, Geusebroek JM, Veenman CJ, Snoek CGM, Smeulders AWM (2006) Robust scene categorization by learning image statistics in context. In: International workshop on semantic learning applications in multimedia, p 105
11. Hauptmann A, Yan R, Lin WH (2007) How many high-level concepts will fill the semantic gap in news video retrieval? In: Sebe N, Worring M (eds) Proceedings of the 6th international conference on content-based image and video retrieval (CIVR 2007). ACM Press, pp 627–634
12. Hiemstra D (1998) A linguistically motivated probabilistic model of information retrieval. In: Proceedings of the 2nd European conference on research and advanced technology for digital libraries (ECDL 1998), pp 569–584
13. Hiemstra D, Rode H, van Os R, Flokstra J (2006) PF/Tijah: text search in an XML database system. In: Proceedings of the 2nd international workshop on open source information retrieval (OSIR 2006), pp 12–17
14. Ho CJ, Chang TH, Lee JC, Hsu JYJ, Chen KT (2009) KissKissBan: a competitive human computation game for image annotation. In: Bennett PN, Chandrasekar R, Chickering M, Ipeirotis PG, Law E, Mityagin A, Provost FJ, von Ahn L (eds) (2009) Proceedings of the ACM SIGKDD workshop on human computation. ACM Press, pp 11–14
15. Joachims T (2002) Optimizing search engines using clickthrough data. In: Proceedings of the 8th annual international ACM SIGKDD conference on knowledge discovery and data mining, pp 133–142
16. Joachims T, Granka L, Pan B, Hembrooke H, Radlinski F, Gay G (2007) Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. ACM Trans Inf Syst 25(2). doi:10.1145/1229179.1229181
17. Joachims T, Li H, Liu TY, Zhai C (2007) Learning to rank for information retrieval (lr4ir 2007). SIGIR Forum 41(2):58–62
18. Kelly D, Teevan J (2003) Implicit feedback for inferring user preference: a bibliography. SIGIR Forum 37(2):18–28
19. Li X, Snoek CGM (2009) Visual categorization with negative examples for free. In: Gao W, Rui Y, Hanjalic A, Xu C, Steinbach EG, El-Saddik A, Zhou MX (eds) Proceedings of the 17th international conference on multimedia. ACM Press, pp 661–664
20. LSCOM Lexicon definitions and annotations version 1.0. Tech. rep., Columbia University (2006)
21. Macdonald C, Ounis I (2009) Usefulness of quality click-through data for training. In: Craswell N, Jones R, Dupret G, Viegas E (eds) Proceedings of the 2009 workshop on Web search click data (WSCD 2009). ACM, New York, pp 75–79
22. Morrison D, Marchand-Maillet S, Bruno E (2009) TagCaptcha: annotating images with CAPTCHAs. In: Bennett PN, Chandrasekar R, Chickering M, Ipeirotis PG, Law E, Mityagin A, Provost FJ, von Ahn L (eds) Proceedings of the ACM SIGKDD workshop on human computation. ACM Press, pp 44–45

23. Palomino MA, Oakes MP, Wuytack T (2009) Automatic extraction of keywords for a multimedia search engine using the chi-square test. In: Proceedings of the 9th Dutch–Belgian information retrieval workshop (DIR 2009), pp 3–10
24. Poblete B, Baeza-Yates RA (2008) Query-sets: using implicit feedback and query patterns to organize Web documents. In: Huai J, Chen R, Hon HW, Liu Y, Ma WY, Tomkins A, Zhang X (eds) Proceedings of the 17th international conference on World Wide Web, pp 41–50
25. Scholer F, Shokouhi M, Billerbeck B, Turpin A (2008) Using clicks as implicit judgments: expectations versus observations. In: Boughanem M, Berrut C, Mothe J, Soulé-Dupuy C (eds) Proceedings of the 30th European conference on IR research, pp 28–39
26. Setz AT, Snoek CGM (2009) Can social tagged images aid concept-based video search? In: Proceedings of the IEEE international conference on multimedia & expo (ICME 2009), pp 1460–1463
27. Sigurbjörnsson B, van Zwol R (2008) Flickr tag recommendation based on collective knowledge. In: Huai J, Chen R, Hon HW, Liu Y, Ma WY, Tomkins A, Zhang X (eds) Proceedings of the 17th international conference on World Wide Web, pp 327–336
28. Smith G, Ashman H (2009) Evaluating implicit judgements from image search interactions. In: Proceedings of the Web science conference: society on-line (WebSci 2009)
29. Snoek CGM, Worring M, van Gemert JC, Geusebroek JM, Smeulders AWM (2004) The challenge problem for automated detection of 101 semantic concepts in multimedia. In: Proceedings of the 14th ACM international conference on multimedia, pp 421–430
30. Tsikrika T, Diou C, de Vries AP, Delopoulos A (2009) Image annotation using clickthrough data. In: Marchand-Maillet S, Kompatsiaris Y (eds) Proceedings of the 8th international conference on content-based image and video retrieval (CIVR 2009). ACM Press
31. Ulges A, Koch M, Schulze C, Breuel T (2008) Learning TRECVID'08 high-level features from YouTube™. In: Proceedings of TRECVID 2008
32. Ulges A, Schulze C, Keysers D, Breuel TM (2008) Identifying relevant frames in weakly labeled videos for training concept detectors. In: Luo J, Guan L, Hanjalic A, Kankanhalli MS, Lee I (eds) Proceedings of the 7th international conference on content-based image and video retrieval (CIVR 2008). ACM Press, pp 9–16
33. Ulges A, Schulze C, Keysers D, Breuel TM (2008) A system that learns to tag videos by watching YouTube. In: Gasteratos A, Vincze M, Tsotsos JK (eds) Proceedings of the 6th international conference of computer vision systems (ICVS 2008). Lecture Notes in Computer Science, vol 5008. Springer, pp 415–424
34. von Ahn L, Blum M, Langford J (2004) Telling humans and computers apart automatically. Commun ACM 47(2):56–60
35. von Ahn L, Dabbish L (2004) Labeling images with a computer game. In: Proceedings of the ACM SIGCHI conference on human factors in computing systems (CHI 2004). ACM Press, pp 319–326
36. von Ahn L, Dabbish L (2008) Designing games with a purpose. Commun ACM 51(8):58–67
37. von Ahn L, Liu R, Blum M (2006) Peekaboom: a game for locating objects in images. In: Grinter RE, Rodden T, Aoki PM, Cutrell E, Jeffries R, Olson GM (eds) Proceedings of the ACM SIGCHI conference on human factors in computing systems (CHI 2006). ACM Press, pp 55–64
38. von Ahn L, Maurer B, Mcmillen C, Abraham D, Blum M (2008) reCAPTCHA: Human-based character recognition via web security measures. Science 321(5895):1465–1468
39. Yang J, Hauptmann AG (2008) (Un)Reliability of video concept detection. In: Luo J, Guan L, Hanjalic A, Kankanhalli MS, Lee I (eds) Proceedings of the 7th international conference on content-based image and video retrieval (CIVR 2008). ACM Press, pp 85–94

Theodora Tsikrika received her PhD in Computer Science from Queen Mary, University of London, UK on the combination of evidence for Web Information Retrieval. In 2007, she joined CWI for a three-year period as a Researcher working on the Vitalas project (FP6). Her research interests include combination of evidence for (multimedia) information retrieval, concept-based multimedia annotation and retrieval, evaluation of (multimedia) information retrieval, Web information retrieval applications, structured document retrieval, and mining and analysis of search interactions logs. Since 2007, she has been involved in the coordination of multimedia retrieval tasks in the INEX and ImageCLEF international evaluation benchmarks. She has served as a Program Committee member for several international conferences and as a reviewer for international journals in the area of (multimedia) Information Retrieval. She is a member of ACM and ACM SIGIR.

Christos Diou graduated from the Department of Electrical and Computer Engineering of the Aristotle University of Thessaloniki, Greece, in 2004 and is currently pursuing a PhD degree in the same department. His research interests lie in the area of concept-based multimedia retrieval and automatic annotation of images and videos. He is a member of the Technical Chamber of Greece and a student member of the IEEE.

In 2005, Mr. Diou was awarded the Greek State Scholarships Foundation Fellowship.

Arjen P. de Vries is a senior researcher at CWI and a part-time full professor in the area of multimedia data spaces at the Technical University of Delft. He received his PhD in Computer Science from the University of Twente in 1999, on the integration of content management in database systems. He is especially interested in the design of database systems that support search in multimedia digital libraries. He has worked on a variety of research topics, including (multimedia) information retrieval, database architecture, query processing, retrieval system evaluation, and ambient intelligence. He has coordinated the TREC and INEX Entity Ranking tracks. In 2004, De Vries and his then PhD student Westerveld received the best paper award in the international conference on image and video retrieval (CIVR), and in 2007, De Vries and his PhD student Cornacchia received the best student paper award in the European conference on Information Retrieval (ECIR). He is currently participating in the EU project PuppyIR (FP7), after recently successfully completing the Vitalas project (FP6).

Anastasios Delopoulos graduated from the Department of Electrical Engineering of the National Technical University of Athens (NTUA) in 1987, received the M.Sc. from the University of Virginia in 1990 and the Ph.D. degree from NTUA in 1993. He is with the Electrical and Computer Engineering Dept. of the Aristotle Univ. of Thessaloniki where he serves as an assistant professor. His research interests lie in the areas of multimedia data understanding and computer vision. He is author of more than 75 journal and conference scientific papers. He has participated in 21 European and National R&D projects related to application of signal, image, video and information processing to entertainment, culture, education and health sectors. Dr. Delopoulos is a member of the Technical Chamber of Greece and IEEE.