IMPROVING VIDEO CONCEPT DETECTION USING UPLOADER MODEL

Usman Niaz and Bernard Merialdo

EURECOM, Campus SophiaTech
450 Route des Chappes, 06410 Biot FRANCE

ABSTRACT

Visual concept detection is a very active field of research, motivated by the increasing amount of digital video available. While most systems focus on the processing of visual features only, in the context of internet videos other metadata is available which may provide useful information. In this paper, we investigate the role of the uploader information, the person who uploaded the video. We propose a simple uploader model which includes some knowledge about the content of videos uploaded by a given user. On the TRECVID 2012 Semantic Indexing benchmark [1], we show that this simple model is able to improve the concept detection score of all the 2012 participants, even the best ones, by only re-ranking the proposed shots. We also present some statistics which show that even though most TRECVID systems are based on visual features only, they provide results which are biased in favor of test videos for which the uploader was present in the development data. This work suggests further research on the use of metadata for visual concept detection, and a different way of organizing benchmark data to assess the visual performance of detectors.

Index Terms— concept detection, uploader bias

1. INTRODUCTION

Despite the recent achievements and the marvels in video content understanding community, industrial video retrieval is still done through the text present in the video metadata namely the video title and the tags associated with the videos. The reasons are the immense number of categories to be indexed, limitations in representing the visual space and the complexity of this representation, and the limitations of learning algorithms among others. Visual and / or multimedia detectors are becoming stronger and naturally more complex as they try to learn all possible variations with which objects appear in a category. Nevertheless, video retrieval takes cues from diverse sources of information other than the mere use of complex visual classifiers like for example the text present in the metadata, interests or preferences of the user retrieving the results, ontological knowledge [2] etc.

The question we are interested in exploring is can the name of the uploader bring any information whatsoever for concept detection in videos? Remember uploaders are the first annotators of the video. The uploader can be either someone who stumbles upon a particular video and decides to share it or can be someone involved with the making of the video. Either way the uploader is aware of the video content and the first and a reliable annotator. We may also wonder how can this uploader's information, if present, be used successfully in order to improve the annotation of videos. We further investigate whether good video concept detection systems are aware of this uploader's information and do they inherently or implicitly use it? The question that then comes to mind naturally is would it be too dangerous to rely on this information and does it harm the essense of true video retrieval? In this paper we try to answer these questions and where ever possible through extensive experimentation.

The state of the art general purpose video concept detection can benefit a lot if a reliable model based on uploader's name is present. Knowing a video is uploaded by a particular user can immensely reduce the search space for indexing or increase assurance or certitude for certain concept(s) if that user's interests are known. To achieve this a good model need to be constructed only from the *uploader name* field of the video metadata. In this paper we present such a model created from the video uploader's name present in the metadata. It is efficient and easy to compute and is only applied to a concept if reliable cues are present. These cues are inferred by judging the consistency of the content uploaded by a user.

We show the effectiveness of the simple and efficient uploader model trained on the metadata belonging to training videos. It improves video concept detection performance for our baseline system which is built using a pool of classifiers based only on visual descriptors. Moreover it improves detection scores for most if not all of the multimedia systems submitted to the light Semantic Indexing task in the TRECVID 2012 campaign [1].

The rest of this paper is organized as follows. Section 2 presents review of some methods to build user models capturing user's preferences, interests and other knowledge, and the use and influence of such models on information retrieval and related tasks. Section 3 presents our motivation and presentation of the uploader model. Section 4 shows the effectiveness of uploader model on challenging TRECVID 2012 dataset. Finally section 5 concludes the paper.

2. RELATED WORK

We get our inspiration from the user modeling in adaptive web systems and personalized information retrieval systems. In these systems user's knowledge and interests are mined to build a model depicting his/her preferences. These systems, like e.g. web search, cater each user's need separately depending on his/her interests like guiding the user to most appropriate content or prioritizing most relevant items returned from the search result [2]. To remain adaptive to user needs they maintain up-to-date user models. The model is updated with the information gathered either explicitly or implicitly through user interaction with the system over time. Parameters used for building such models are usually user's knowledge, interests, goals, background, and individual traits [2].

Hadjouni et al. [3] present a personalized data retrieval approach based on a multi-dimensional end user model constructed iteratively. Numerical values for a web entity are calculated based on user interest model which is built implicitly from the duration and frequency of visits to the visited links by that user. These values are used to calculate the *semantic distance* between two entities to find the most relevant ones for that user. For unknown users initially the closest known user's profile is used while a real-time profile is being constructed. Larson et al. [4] use the ID of the uploader contained in the metadata for refining geo-tagging in the 2010 MediaEval *Placing* task.

Zhang et al. [5] model user preferences for personalized retrieval of sports videos. Classical text based retrieval is used to return initial results which are used to capture user's interests. The user browses through the desireable videos from those returned and that *click-through* data is used to build a semantic and a visual feature based user preference model. The semantic model is built using the annotation fields of the clicked video clips. For an incomplete query, where all the fields are not mentioned by the user, relevant results are randomly presented. Important insights are obtained from the fields of the clicked through clips. More specifically the mentioned and unmentioned fields are used to obtain weights for each field of the clicked clip and the user's query. This models the user's preference for the current query. Finally a *semantic score* is calculated between each returned clip and the clips clicked by user using those weights in the preference model. This semantic score is higher for a desireable video. To model the user preferences based on visual information SVM classifiers are used on separate sets of visual features extracted from the video clips to classify them into satisfied and unsatisfied. Clips clicked by the user are considered positive while others upto the last clicked clip are taken as negative examples. Each classifier gives a probabilistic output score which are combined linearly weighted by the normalized inverse variance of the features. The semantic and visual models can be used independently or with a weighted combination to re-rank the retrieved results according to the user's preferences.

Xu et al. [6] use an authority score for each user to remove tags that may be spam while identifying the most appropriate tags. The authority score measures the average quality of the user's tags. This quality increases if the tags provided by the user coincide with majority of the tags given by other users. The initial weight is set the same for each user which is then updated iteratively over time based on that user's tagging information. In this authority score higher weights are given to the users who originally tagged the content (potential uploaders). In the end tags assigned by more authoritative users, among other criteria, are considered accurate for the web content i.e. the most authoritative user labels the content.

Bueno and David [7] build an explicit individual user model (rather than a user-group model) for representing user's activitites and interests for personalized information retrieval. A user's goal is represented as a query in natural language which is also called the user objective. The user model contains the user's evaluations (*ok*, *known*, *?*, *wrong*) for all the evaluated documents for each objective. All the documents in their documents retrieval system are parameterized mainly with keywords and also with author, year or even name of the journal. For each of these parameters the value of the evaluations is incremented by one after each evaluation of the document by the user. They use Naive Bayes to calculate the degree of relevance of an object to the present objective for a user using those values.

Sugiyama et al. [8] propose to adapt search results according to user's information needs. They allow a fine grained search for each user by updating a user's profile model capturing changes in his/her preferences. Probabilistic user profiles are built from their browsing history containing a long term or persistent component updated over time and another short term or ephemeral component reflecting the current day's activity. These components are weighted in order to highlight the importance of recent events over the others or vice versa. Web pages returned form the search results are matched with the user profile to determine the *similarity* and relevant results are then presented to the user. We built users model based on the content that he/she has uploaded. In our case the fewer the number of different concepts a user has uploaded the greater is our confidence in the presence of those concepts in other videos uploaded by the same user. Relationsips between users interests are not exploited in our approach.

These adaptive web systems and personalized information retrieval systems dictate that if a user has so and so interests he/she should be presented with such content. We adapt this notion in the context of internet video retrieval to build an uploader model depicting that if a user has uploaded such videos the probability of the presence of so and so concepts is higher.

3. UPLOADER MODEL

Before calculating a reliable uploader's model let us look at some statistics acquired from the TRECVID 2012 [1] data.

	Videos	Videos with Uploaders	Different uploaders
Development	19,701	19,331 (98.1%)	4,415
Test	8,263	8,073 (97.7%)	2,505

Table 1. Intra-collection statistics for TRECVID 2012 development and test datasets

3.1. How Reliable is Uploader's Information

The TRECVID 2012 collection consists of 800 hours of internet videos from which 600 hours or approximately 19,000 videos are present in the training set and the remaining approximately 8000 videos are used for testing purposes. We first present some intra-collection statistics in the table 1, which show that for most of the videos we have the uploader's name present in the metadata and also show the number of different uploaders which is significant.

We next determine the percentage of videos in the test for which the uploader is present in the development set. Table 2 shows that this percentage is significant which compelled us to use the uploader's information for improving the results. Further we look at the distribution of uploaders in figure 1 and conclude that most of the uploaders have posted adequate number of videos.

We analyze if using the uploader's name result in any sort of information gain. For this purpose we calculate the entropy of the 346 concepts provided with the TRECVID 2012 dataset, figure 2. The entropy of less frequent concepts like *Yasser_Arafat* is close to zero meaning that guessing absence of those concepts for any video is almost always right. On the other hand for the more frequent concept the entropy or the measure of uncertainty increases. For example the concept *Trees* with entropy close to 1 is present in almost half of the test videos. From figure 2 we see that using uploader's name with the concept name results in decreased entropy for all the concepts. That is knowing the uploader's name reduces the uncertainty for each concept.

Test	**Total**	**With dev uploader**	
Videos	8,263	6,914	83.7%
Shots	145,634	118,845	81.6%

Table 2. Inter-collection statistics for the TRECVID 2012 test dataset

We thus try to use the uploader's information in an efficient way to improve video concept detection results. The next subsection describes the model in detail which is calculated for each concept from the training data and is applied only to the concepts for which reliable information is present.

3.2. Uploader Model

An uploader is likely to upload videos containing similar content or related/belonging to similar concept. For example if a user runs a video blog about monuments in a certain city then almost all videos uploaded by that user will contain concepts like *sky* or *outdoor*. This information thus increases our confidence in the predictions of the concepts *sky* and *outdoor* if the test video is uploaded by the same user.

The uploader model simply calculates the ratio of video shots uploaded by the uploader for each concept from the training data and modifies the output score of each new video shot if that video is uploaded by the same person. This uploader bias allows us to rerank the retrieval results. For each concept we calculate the probability of concept given uploader as:

$$p(c/u) = \frac{W_u^c}{|V_u|}$$

where V_u is the set of videos uploaded by uploader 'u' and W_u^c is the weightage of videos uploaded by uploader 'u' for

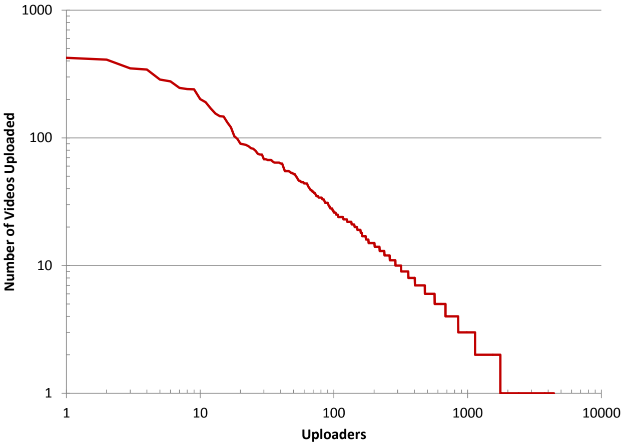

Fig. 1. Uploader distribution

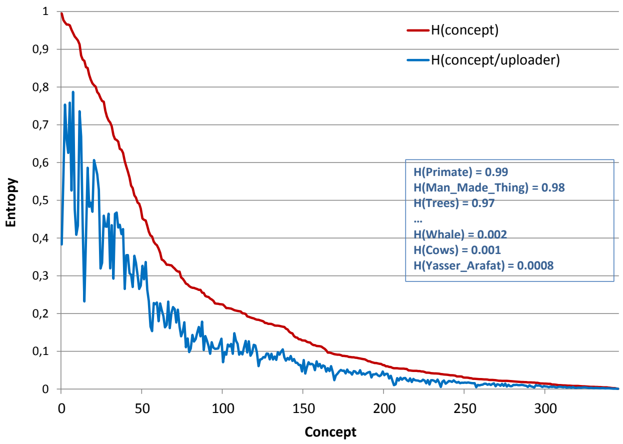

Fig. 2. Entropy of the concepts vs Entropy of the concepts given uploader's name

the concept 'c'. This quantity:

$$W_u^c = \sum_{v \in V_u} \frac{|s \in v, s.t. s = c|}{|s \in v|}$$

is the sum of ratios of the number of shots labeled with concept 'c' to the total shots in that video for all the videos uploaded by 'u'.

We also calculate average uploader's probability for each concept as:

$$p(c) = \frac{W^c}{|V|}$$

where W^c is the total weightage of all the videos uploaded for concept 'c', given by:

$$W^c = \sum_u W_u^c$$

and V is the number of videos, or

$$V = \sum_u V_u$$

.

This model is computed on the training data separately for each concept. To apply the uploader's model to the test videos we calculate the coefficient α as:

$$\alpha(c, u) = \max\left(\frac{p(c/u) - p(c)}{p(c/u) + p(c)}, 0\right)$$

Concepts with Uploader Model	Concepts without Uploader Model
Adult	Asian_People
Airplane_Flying	Building
Animal	Bus
Bicycling	Cheering
Boat_Ship	Cityscape
Car	Classroom
Dancing	Computer_Or_Television_Screens
Dark-skinned_People	Computers
Female_Person	Demonstration_Or_Protest
Flowers	Doorway
Indoor	Explosion_Fire
Indoor_Sports_Venue	Female-Human-Face-Closeup
Infants	Ground_Vehicles
Instrumental_Musician	Hand
Male_Person	Helicopter_Hovering
News_Studio	Landscape
Old_People	Military_Base
Running	Mountain
Singing	Nighttime
Sitting_Down	Plant
Stadium	Road
Swimming	Scene_Text
Telephones	Vehicle
Throwing	Walking
Waterscape_Waterfront	Walking_Running

Table 3. Concepts to which the uploader model was applied after our visual only run

The uploader model is applied on top of a video retrieval system as a re-ranking mechanism. The concept score of each shot $p(c|s)$ from the detection system is updated as follows:

$$p_u(c|s) = p(c|s) * (1 + \alpha(c, u))$$

The model is applied only to the concepts for which a significant improvement is acheived in the development set. In this case it was applied to 25 concepts out of a total 50 concepts shown in the table 3.

4. EXPERIMENTS

All the experiments are performed on the TRECVID 2012 dataset [9, 1]. As the TRECVID giudelines dictate we have used almost three quarters of the dataset for development purposes and the fourth quarter is used for test.

4.1. OUR Light SIN Results

We have first applied and validated the uploader model on our submission to the TRECVID light Semantic INdexing (SIN) [1] task which contains detecting 50 concepts. The model was applied on top of our visual run containing a set of 10 visual features. Details of that system can be found in [10]. The uploader model was only applied to 25 out of 50 concepts for which there was improvement shown on the development set. Of the total 50 concepts only 16 were evaluated by NIST for the test set and of those 16, 10 were present in the set of concepts to which uploader model was applied. Figure 3 shows the Average Precision (AP) scores of the 16 evaluated concepts. The uploader model is applied to the 10 concepts shaded in blue. NIST calculates the AP scores for the first 2000 shots returned for each concept [9].

Using the uploader bias is shown to be effective with a considerable increase in concept detection performance over the previous run even though it is applied to only 10 concepts out of 16. The model improves average precision for almost every concept to which it is applied except for *Male_Person* where there is a negligible 0.8% drop in the already high score. AP score for the poorly recognized concept *Sitting_Down* drops from 0.0015 to 0.0008. Other than that AP score for every concept is increased with the uploader model. Some considerable improvements in the score are 158% for the concept *Stadium*, 95% for *Instrumental_Musician* and 60% for the concept *Airplane_Flying* among others, with an average improvement of 21% over our Visual-only system. Furthermore using uploader's information to detect concepts proves beneficial as it increases score even when the visual descriptors failed to retrieve the concept *Bicycling*.

4.2. All TRECVID SIN Runs

To further evaluate the effectiveness of the uploader model we try to apply it to runs submitted by all other participants in

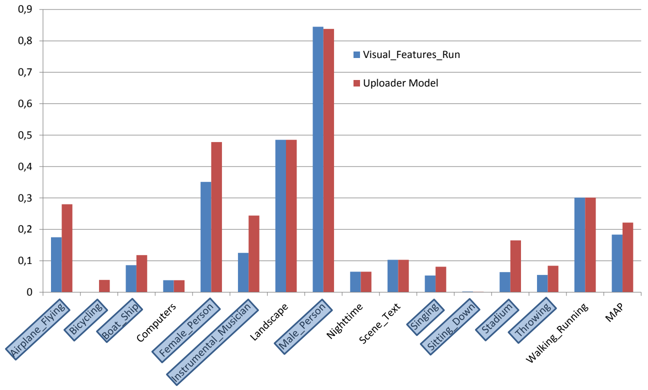

Fig. 3. Improving *our* SIN runs: Average Precision score of the evaluated concepts after applying the uploader model (blue boxes) and the Mean Average Precision of both runs.

the TRECVID 2012 light SIN task. Unfortunately we do not have the actual classification scores for each shot from the participants but only the ordered list of best 2000 shots for each concept. This information is returned to each participant after all the results are evaluated at the end of the TRECVID campaign. We devise a method to find an artificial score for each shot based on the inverse of its rank. As the list of shots is ordered we have the first short scoring the highest. So the socre can be calculated as:

$$score = \frac{100}{100 + rank}$$

where rank is just the numerical index of the shot.

As alternatives to this artificial score we use the score of one of our runs corresponding to the 2000 shots for each concept and for each run in the TRECVID SIN task. We also use the average score of all of our runs corresponding to the 2000 shots returned by each participant for each concept.

Another limitation that we face here is that we do not have best shots after 2000. Since each participant returns only the first 2000 shots we are bound to work with only those. So instead of the whole list of shots we re-rank only the best 2000 for each concept using the uploader model. The re-ranking results are shown in the figure 4 showing improvements for almost all the runs with the three kinds of scores. There are a total of 91 systems or runs where each team was allowed to submit a maximum of 4 runs. The runs are sorted in the figure 4 with performance along the x-axes and we can see that the low performing ones benefit more from the uploader based re-ranking. This is true for the three kinds of scores used with the Average Score giving maximum improvement for most runs, while using Artificial Scores in the uploader model improves performance of all the runs submitted to the TRECVID 2012 light SIN task.

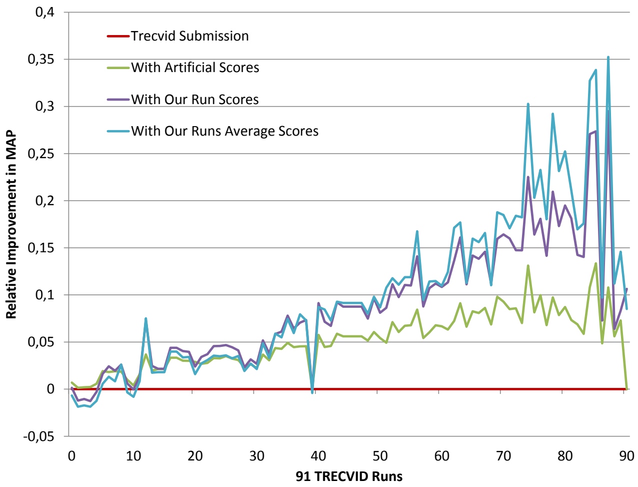

Fig. 4. Improving (all) TRECVID SIN runs: Improvement in Mean Average Precision of the 91 systems over the baseline (TRECVID submission) with the uploader model using three kinds of scores. Artificial score used for uploader model improves all the submitted runs.

The overall improvement for all the Light SIN task runs is shown in the table 4. Remember the uploader model is applied to only 10 out of the 16 evaluated concepts as before for each of the runs.

4.3. Performance of Runs vs Uploaders Information

We try to find if there is some correlation between high performance runs and the uploader information, i.e. do good video retrieval systems leverage from the uploader's imformation present in the datasets. This can be easily checked by seeing if there are any uploaders from the training set present in the returned ranked lists. More specifically we find the percentage of uploaders from training present in the top ranked shots returned from the system. For this purpose we select five best performing teams in the TRECVID light SIN task and take the results for the best system for each of those. Figure 5 shows the average percentage of uploader in the results of the five systems ranked against the number of shots. The systems are also ranked by their performance on the test set, i.e. Run 1 is the best one, then comes Run 2 and so on in the figure 5. The percentage of uploaders decreases with the total number of shots and it is evident that the best systems do actually benefit from implicit use of uploaders information.

No change	Inv rank	OUR run scores	OUR run average score
19.14%	19.89%	20.29%	20.32%

Table 4. Improving all TRECVID Light SIN results: Average percentage improvement over the 91 systems for the three kinds of scores.

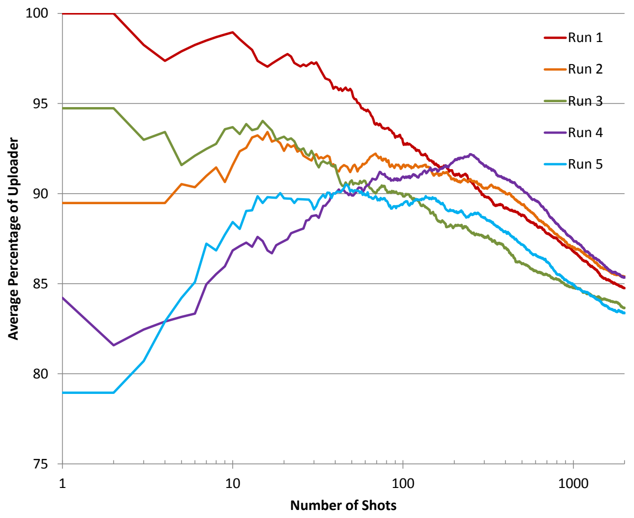

Fig. 5. Percentage of uploaders in the results by best systems from five teams

5. CONCLUSIONS

Information based on uploader of the video tells a lot about the video as a person tends to upload similar videos. This phenomenon is refected in the results with the use of a simple reranking model based on videos' uploader. Not only were we able to improve our own submission to the TRECVID 2012 Light Semantic Indexing Task but we also showed improvements on almost all of the submitted run results from the other participants. We also showed that there is some correlation between better runs and uploaders and concluded that better TRECVID performance demands the use of information based on uploader's identity.

The use of uploader model on video concept detection systems has potential for improvements and requires future work. Consequently we would like to enhance the uploader model to find the best model for maximum improvement over the visual features based systems. Also there is a need to find a method to update scores of different kind. So far we have applied the uploader model to artificial scores and corresponding scores from our own runs. This does not mean that this model is optimal if we had the actual scores from the TRECVID participant systems. Since a classifier can output scores of different kinds and those scores can be fused in a variety of ways. We would further like to work on the model to incorporate the case where more than 2000 shots (or complete results) are presented for re-ranking. As one concept is usually uploaded by multiple users we can try to find correlations between the users to bring more information to the uploader model. Users can be clustered in groups according to their interests and the confidence in the prediction of a test video uploaded by a group member can be altered according to the group interest.

We have seen from the results that there seem to be a correlation between the uploader information and visual detectors. Certain good systems based on visual only descriptors are sometimes using this correlation and it would be misleading to grant their good performance only to the visual classifiers. We argue that to assess the true impact / power of systems based on visual descriptors the uploaders in the training and test collections should be disjoint. In other words a disjoint collection will minimize the implicit reliance of visual only systems on the use of uploaders' information.

6. REFERENCES

[1] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B Shaw, W. Kraaij, A. F. Smeaton, and G. Quenot, "Trecvid 2012 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proceedings of TRECVID 2012*. NIST, USA, 2012.

[2] P. Brusilovsky and E. Millán, "The adaptive web," chapter User models for adaptive hypermedia and adaptive educational systems, pp. 3–53. Springer-Verlag, 2007.

[3] M. Hadjouni, M.R. Haddad, H. Baazaoui, M.A. Aufaure, and H. Ben Ghezala, "Personalized information retrieval approach," in *WISM*, 2009.

[4] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, and G. J. F. Jones, "Automatic tagging and geotagging in video collections and communities," 2011, ICMR.

[5] Y. Zhang, C. Xu, X. Zhang, and H. Lu, "Personalized retrieval of sports video based on multi-modal analysis and user preference acquisition," *Multimedia Tools Appl.*, vol. 44, no. 2, pp. 305–330, Sept. 2009.

[6] Z. Xu, Y. Fu, J. Mao, and D. Su, "Towards the semantic web: Collaborative tag suggestions," in *Collaborative Web Tagging Workshop at WWW*, 2006.

[7] D. Bueno and A. David, "Metiore: A personalized information retrieval system.," in *User Modeling*, M. Bauer, P. J. Gmytrasiewicz, and J. Vassileva, Eds. 2001, vol. 2109, pp. 168–177, Springer.

[8] K. Sugiyama, K. Hatano, and M. Yoshikawa, "Adaptive web search based on user profile constructed without any effort from users," in *WWW*, 2004.

[9] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *MIR '06*, 2006, pp. 321–330.

[10] U. Niaz, M. Redi, C. Tanase, and B. Merialdo, "EURECOM at TrecVid 2012: The light semantic indexing task," in *TRECVID*, 2012.