Uploader models for Video Concept Detection

Bernard Merialdo, Usman Niaz
Multimedia Communications Department
EURECOM
Sophia Antipolis, FRANCE
{FirstName.LastName}@eurecom.fr

***Abstract*—In video indexing, it has been noticed that a simple uploader model was able to improve the MAP of concept detection in the TRECVID Semantic Concept Indexing (SIN) task. In this paper, we explore this idea further by comparing different types of uploader models and different types of score/rank distribution. We evaluate the performance of these combinations on the best SIN 2012 runs, and explore the impact of their parameters. We observe that the improvement is generally lower for the best runs than for the weaker runs. We also observe that tuning the models for each concept independently produces a much more significant improvement.**

Keywords—Multimedia Indexing, User model, TRECVID

I. INTRODUCTION

With the ever increasing amount of digital video documents on sites such as YouTube, Dailymotion or others, the task of automatically indexing the content of those documents has become a major research issue. A basic component is to be able to detect the occurrence of a concept (objet, action, scene) inside a video. While most approaches have focused on developing computer vision techniques, the recent explosion of user generated content and social networks has brought a lot of attention on users information. Since systems based on visual recognition are far from perfect, they provide results with a large degree of uncertainty. Combining the visual information with information extracted from the relations between the video and the users may improve the capacity of systems to better identify the content of a video.

In this paper, we explore the use of uploader information in the task of video concept detection. The uploader is the user who has uploaded the video on an internet site. Given that each user has specific interests, it is likely that knowing the uploader information gives a prior knowledge about the possible concepts which can be found in a video. This information can be usefully combined with the visual information to improve the accuracy of concept detection.

We focus on the experiments conducted within the TRECVID evaluation campaign [1], one of the major benchmarks for the comparison of video indexing systems. In the TRECVID Semantic Indexing (SIN) task, a list of video shots S is provided, together with a list of semantic concepts C. The goal is to detect in which shots each concept appears. For each concept, the result is a ranked list of the 2000 shots where the concept is most likely to appear. Development and test data are provided to the participants.

The generic approach to the SIN task is to extract from each shot s one or a set of features F(s) (mostly video, but sometimes audio) and use the development data to train classifiers. The output of the classifier is a score Score(s,c), with $s \in S$, $c \in C$ which is higher when the classifier is more confident about the occurrence of the concept within the shot.

Traditionally, the score is computed from the features only, as a combination of the scores over the individual features:

$$Score(s,c) = VScore(F(s),c)$$

As each shot s originates from a video v=v(s) which has been uploaded by a user u=u(v(s)), we can describe the shot not only by its features, but also by the uploader id:

$$Score(s,c) = Score(F(s),u,c)$$

By analogy with probabilities, and assuming independence between the feature spaces and the uploader, we can introduce the uploader information as a multiplicative factor $\alpha(c|u)$ applied to the traditional score Score(s,c). In this paper, we thus defined a combined model by:

$$Score(s,c) = VScore(F(s),c) * \alpha(c|u)$$

II. RELEVANT WORK

User modeling has been a long standing line of research in information management, especially 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 use these preferences to guide each user to the most appropriate content or to prioritize the most relevant items returned from the search result [10]. To remain adaptive to user needs, 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 [10].

Zhang et al. [11] 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 desirable 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 desirable video. To

978-1-4799-3990-9/14/$31.00 ©2014 IEEE

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. [12] 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 [13] build an explicit individual user model (rather than a user-group model) for representing user's activities 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. [14] 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.

Recently, user information has started being introduced in models used for the prediction of multimedia content. In the MediaEval campaign, Rouvier et al. [8] have used this information to substantially improve the accuracy of their system. Similar attempt by. Xu et al. [9] has not been as successful, probably because the relationships for the combination between user information and multimedia analysis is not yet completely understood, neither is the most efficient representation for a user model. This has motivated us for a more extensive exploration of various user models and combination with score distributions on ranked lists.

III. SCORE RANK DISTRIBUTIONS

The TRECVID submissions do not contain any score information, but just a ranked list of shots. We therefore transform the rank into a score value according to a predefined Score Rank Distribution. In this paper, we experiment with the following Score Rank Distributions, that are classical models proposed in the literature on Information Retrieval systems.

*A. **Reciprocal Rank model (RR)***

The Reciprocal Rank is based on the inverse of the rank. It has been shown to be efficient for the fusion of ranked list in information retrieval [6]. In order to balance the importance of the first elements in the list, an offset a is added to the rank.

$$VScore(x) = \frac{1}{a + Rank(x)}$$

*B. **Borda Count model (BC)***

The Borda Count [3] is just a linear function of the rank. We normalize it, so as to have a value between 0 and 1, and again include a tuning parameter.

$$VScore(x) = 1 - a * \frac{Rank(x)}{N}$$

*C. **Logistic model (LM)***

The Logistic Model [3] uses a logistic function to derive the score from the rank. While this model uses 2 tuning parameters, we arbitrarily fixed the value of the ratio between these two parameters, so as to reduce the search space.

$$VScore(x) = \frac{1}{1 + e^{a*Rank(x)+b}}$$

*D. **Informetric Distribution model (ID)***

The Informetric Distribution [3] is derived from the studies of the frequency versus rank distribution, of which the most famous example is the Zipf law. The tuning parameter is here the exponent of the inverse of the rank, and varies with values around and close to 1.

$$VScore(x) = \frac{1}{Rank(x)^a}$$

IV. UPLOADER MODELS

As previously explained, we define the uploader model as a multiplicative factor $\alpha(\mathrm{c}|\mathrm{u})$ to the score of a shot for a given concept. We experiment with several models, build from the combination of the probability of the concept given the user $p(c|u)$, the average probability of the concept $\bar{p}(\mathrm{c})$, the number $n(u)$ of videos uploaded by u, and a coefficient β which is optimized in the experiments.

We use the following models:

- Boosting by the probability of the concept
  UM1: $\alpha(\mathrm{c}|\mathrm{u}) = 1 + \beta * p(c|u)$
- Smoothed probability with fixed weight
  UM2: $\alpha(\mathrm{c}|\mathrm{u}) = (1-\beta) * \bar{p}(\mathrm{c}) + \beta * p(c|u)$
- Smoothed probability with importance weight
  UM3: $\alpha(\mathrm{c}|\mathrm{u}) = \frac{\beta}{\beta+n(u)} * \bar{p}(\mathrm{c}) + \frac{n(u)}{\beta+n(u)} * p(c|u)$
- Adhoc model
  UM4: $\alpha(\mathrm{c}|\mathrm{u}) = 1 + \beta * \frac{p(c|u)-\bar{p}(\mathrm{c})}{p(c|u)+\bar{p}(\mathrm{c})}$
- Positive Adhoc model (used only when $p(c|u) > \bar{p}(\mathrm{c})$)
  UM5: $\alpha(\mathrm{c}|\mathrm{u}) = 1 + \beta * \frac{p(c|u)-\bar{p}(\mathrm{c})}{p(c|u)+\bar{p}(\mathrm{c})}$

For consistency, when applying the uploader model to a shot of a video in the test collection, if the uploader does not appear in the development collection, or if the uploader information is missing, the multiplicative factor is the value corresponding to the average probability of the concept. It should also be noted that the uploader model provides the same value for all the shots inside a video.

V. EXPERIMENTS

A. *TRECVID SIN Data*

The TRECVID SIN 2012 development data contains 400,289 shots, extracted from 19,701 videos (about 400 hours) downloaded from the Internet Video Archive [2]. The test data contains 145,634 shots, extracted from 8,263 videos (about 200 hours). The Full task identifies a set of 346 semantic concepts, out of which 46 have been selected for evaluation. The development data has been partially manually annotated using a collaborative platform [5].

If we observe the development and test collections, we notice that the uploader information is present for most of the videos. TABLE I. shows the statistics for each set.

TABLE I. STATISTICS OF VIDEOS WITH UPLOADER

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

Furthermore, if we search how many videos and shots of the test collection have an uploader who already uploaded videos in the development collection, we find the following figures, as shown in TABLE II.

TABLE II. STATISTICS OF TEST VIDEOS WITH DEV. UPLOADER

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

A large percentage of the videos and shots of the test data have an uploader who is present in the development data. This motivates the use of an uploader model created from the development data and used to improve the prediction on the test data.

Our experiments are performed on the best runs of the best 6 teams in the Full task, which we identify as Run1, Run2,…Run6. The Mean Average Precision of these runs are given in TABLE III.

TABLE III. MAP OF THE BEST RUNS OF THE 6 BEST TEAMS – FULL TASK

Average	**Run1**	**Run2**	**Run3**	**Run4**	**Run5**	**Run6**
25.97	32.10	29.68	26.92	23.79	22.63	20.69

Note that for these runs, only the ranked list of shots from the test data is available. This prevents us from tuning the parameters of the models on validation data, therefore we perform the experiments by directly tuning the parameters on the test data. As such, the results that we obtain in the experiments to follow should be considered as upper bounds of the improvement that can be brought by the uploader models. However, because the results show a low sensitivity to the actual values of the parameters, we are confident that a substantial share of this improvement could be obtained by proper tuning on validation data.

B. *Score-Uploader combination*

In this experiment, we apply the uploader models on the various score distributions, with different values of the coefficients used in the definition of the model. For each combination of values of these coefficients, we compute the average of the mean average precision for all 6 runs, and we report the maximum of this average in the following table.

TABLE IV. AVERAGE MAP OF 6 RUNS WITH VARIOUS SCORE AND UPLOADER MODEL COMBINATIONS

	UM1	**UM2**	**UM3**	**UM4**	**UM5**
RR	27.53	27.67	27.69	27.40	27.43
BC	27.47	27.17	26.76	27.38	27.47
LM	27.04	27.35	27.33	27.25	27.06
ID	27.66	27.80	**27.94**	27.46	27.46

TABLE IV. shows that the uploader model is always beneficial to the total score. The Informetric Distribution model ID provides the best results for most of the uploader models, while the Smoothed probability with importance weight UM3 is the best uploader model.

TABLE V. below shows for each run the initial performance, the performance obtained after application of the UM3+ID combination, the relative improvement, and NC indicates the number of concepts for which the uploader model has a positive impact. It can be observed that the best runs get less benefit from the uploader model than the worse runs: Run6 gets almost 13% improvement, while Run1 gets only 3.2%. It also shows that most of the concepts benefit from the uploader model.

TABLE V. COMPARISON OF INITIAL AND BEST MAP FOR ALL 6 RUNS

	Avg	**Run1**	**Run2**	**Run3**	**Run4**	**Run5**	**Run6**
Initial	25.97	32.10	29.68	26.91	23.79	22.63	20.69
UM3 +ID	27.94	33.14	31.34	28.74	25.77	25.29	23.36
Imp.	7.60%	3.24%	5.57%	6.80%	8.31%	11.76%	12.92%
NC		30	42	39	41	45	44

In **Error! Reference source not found.**, we provide the details of the improvement for each evaluated concept and each run, ordered by decreasing values of the average improvement. The concepts which occur at the top of the table are those for which there is a stronger correlation between the concept appearance and the uploader, therefore the impact of the uploader model is greater. "*Motorcycle, Girl, Bicycling, Baby, Basketball*" are examples of concepts with high correlation. The concepts at the bottom of the table show poor correlation between their occurrence and the uploader. "*Roadway_Junction, Clearing, Landscape, Bridges, Scene_Text*" are example of concepts with low correlation.

C. *Uploader concept correlation*

We performed another experiment to visualize the impact of uploader information on the shots selected by the various runs. We looked at the first N shots provided by the runs for a concept and computed the percentage of shots coming from videos by an uploader known in the development set. The plots of this percentage for values of N ranging from 1 (the best shot) to 2000 (the maximum size provided by TRECVID runs) are shown in Fig. 1. With the exception of Run1, which is an outlier, there is a strong correlation between the value of this percentage and the quality of the run, and higher percentage is indicative of better run.

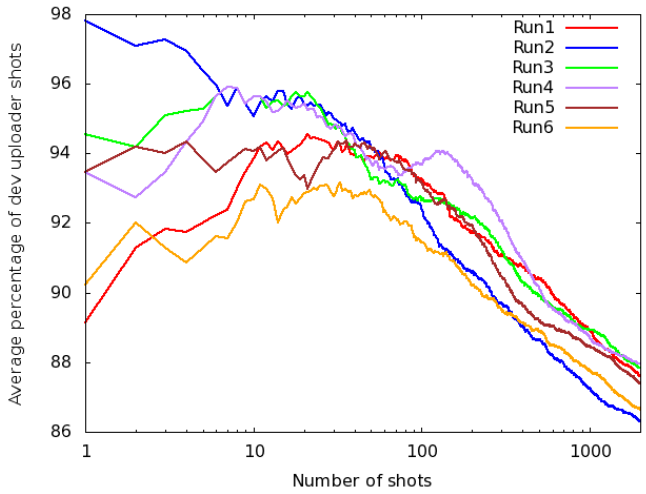

Fig. 1. Percentage of shots with dev uploader in the 6 runs

It is also clear that all runs greatly benefit of uploader information: the average percentage of dev. uploader shots is 81.6% for the test collection, while it is higher than 90% if we consider the first shots of the runs (which provide most of the contribution to the Average Precision). This shows that the runs are actually taking advantage of the visual similarity between videos from the same uploader in the development and test collection.

VI. CONCLUSION

In this paper, we have investigated the use of uploader information to improve the detection of semantic concepts in videos. We have proposed a set of score-rank distributions, and a set of uploader models. The different combinations have been tested on some of the best runs of the TRECVID SIN 2012 benchmark. All combinations show an improvement in the global performance, while UM3+ID seems to be consistently ahead. This improvement is generally lower for the best runs, which suggest that the best runs make a better usage of the inherent correlation between uploaders and the video content they upload.

Several tracks for improving over this approach are possible. For example, many users have uploaded only few videos, which makes the concept statistics quite weak. Grouping similar users together may provide more reliable prediction. Also, we only considered the effect of uploader information as a multiplicative factor to the visual score, while other types of combinations may be more efficient.

We expect that these results are encouraging enough to motivate more research to find the most efficient user models and the best combination mechanisms to improve video concept detection.

VII. REFERENCES

[1] Smeaton, A. F., Over, P., and Kraaij, W. 2006. Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval* (Santa Barbara, California, USA, October 26 - 27, 2006). MIR '06. ACM Press, New York, NY, 321-330.

[2] Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Barbara Shaw, Wessel Kraaij, Alan F. Smeaton, Georges Quenot. 2012. TRECVID 2012 -- An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In Proceedings of TRECVID 2012, NIST, USA.

[3] Wu, Shengli. 2012. *Data Fusion in Information Retrieval*. Springer, ISBN 978-3-642-28865-4

[4] Evangelos Kanoulas, Keshi Dai, Virgil Pavlu, and Javed A. Aslam. 2010. Score distribution models: assumptions, intuition, and robustness to score manipulation. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '10). ACM, New York, NY, USA, 242-249.

[5] Ayache, Stephane and Quenot, Georges. Video Corpus Annotation using Active Learning, European Conference on Information Retrieval, ECIR 2008, Glasgow, Scotland, pp 187-198.

[6] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*(SIGIR '09). ACM, New York, NY, USA, 758-759.

[7] Tin Kam Ho; Hull, J.J.; Srihari, S.N.,1992. On multiple classifier systems for pattern recognition, 11th IAPR International Conference on Pattern Recognition, pp.84,87, 30 Aug-3 Sep 1992

[8] Mickael Rouvier, Georges Linares. 2011. LIA @ MediaEval 2011 : Compact Representation of Heterogeneous Descriptors for Video Genre Classification. *MediaEval 2011 Workshop*, Pisa, Italy

[9] Peng Xu, Yangyang Shi, Martha A. Larson. 2012. TUD at MediaEval 2012 genre tagging task: Multi-modality video categorization with one-vs-all classifiers. *MediaEval 2012 Workshop*, Pisa, Italy

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

[11] Y. Zhang, C. Xu, X. Zhang, H. Lu. 2009. 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.

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

[13] D. Bueno, A. David. 2001. 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.

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