Tag Suggestion on YouTube by Personalizing Content-based Auto-Annotation

Dominik Henter
University of Kaiserslautern
D-67663 Kaiserslautern,
Germany
d_henter@cs.uni-kl.de

Damian Borth
University of Kaiserslautern
D-67663 Kaiserslautern,
Germany
d_borth@cs.uni-kl.de

Adrian Ulges
German Research Center for
Artificial Intelligence (DFKI)
D-67663 Kaiserslautern,
Germany
adrian.ulges@dfki.de

ABSTRACT

We address the challenge of tag recommendation for web video clips on portals such as YouTube. In a quantitative study on 23,000 YouTube videos, we first evaluate different tag suggestion strategies employing user profiling (using tags from the user's upload history) as well as social signals (the channels a user subscribed to) and content analysis. Our results confirm earlier findings that – at least when employing users' original tags as ground truth – a history-based approach outperforms other techniques.

Second, we suggest a novel approach that integrates the strengths of history-based tag suggestion with a content matching crowd-sourced from a large repository of user generated videos. Our approach performs a visual similarity matching and merges neighbors found in a large-scale reference dataset of user-tagged content with others from the user's personal history. This way, signals gained by crowdsourcing can help to disambiguate tag suggestions, for example in cases of heterogeneous user interest profiles or nonexisting user history. Our quantitative experiments indicate that such a *personalized tag transfer* gives strong improvements over a standard content matching, and moderate ones over a content-free history-based ranking.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

Tag Suggestion, Content Analysis, Personalization, Social

1. INTRODUCTION

Web video platforms have experienced an immense growth over the last years in terms of content being broadcasted and their user communities' interaction. This can be seen on YouTube - the market leader in this area - which alone handles 72 hours of video uploads every minute, serves 3 billion hours of video a month and is visited by 800 million unique users each month with half of them taking one of the available social actions, like commenting other videos or sharing videos with friends [1].

To guide users to the content they are interested in, web video platforms rely on textual annotations such as titles and tags. These form the driving force for many applications like search, recommendation, browsing, and advertising. While this meta-data – as added by the video owner manually during the uploading process – is vital for the before mentioned services, it is extremely sparse as illustrated in Figure 1: Lots of YouTube videos are weakly labeled and address only a limited audience [2]. Correspondingly, the majority (99%) of video views on YouTube are generated by a few (30%) highly popular videos [25].

To make sparsely annotated videos more accessible, tag recommendation – the automatic suggestion of potential tags for a clip – can help. Often, such systems provide a set of tags to annotate a user's content alongside his own tags. Additionally, they might be used to automatically deduce tags for videos [21] or images [1] that were not tagged by their original uploader.

Tag suggestion can employ different strategies and infor-

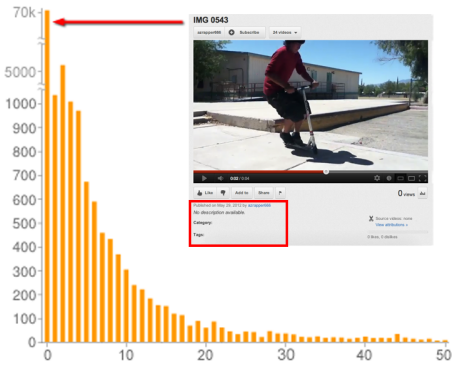

Figure 1: An evaluation [2] showing how many videos are annotated by how many tags. As seen the majority of videos do no have any tags assigned or are weakly annotated. *Top right*: A video as seen on YouTube with no title nor tags assigned (red box).

CrowdMM'12, October 29, 2012, Nara, Japan.

[1] http://www.youtube.com/t/press_statistics
[2] http://www.dfki.uni-kl.de/~borth/blank_videos

mation sources, including **(1)** personalization signals from the user's upload history ("suggest tags that the user has used before") as well as **(2)** social signals ("suggest tags that the user's friends / subscribed interest groups have used") or **(3)** content-based ones ("infer tags from the content of a video"). These different information sources come with different limitations: While history-based tagging is bound to fail in case of topical shifts, content-based tagging is limited as tags are motivated by lots of other factors that are difficult to infer from the content (think of a user's personal background and affect) [1].

Therefore our main contribution is to show that crowd-sourced signals can help when integrated into a tag suggestion system. To do so, we first evaluate common tag recommendations strategies on a large-scale dataset of YouTube clips where the original user-generated tags are used as ground truth. Here, our results confirm earlier findings [14] that a history-based suggestion of tags offers a simple and strong strategy. Second, as our focus moves towards crowd-sourcing, we introduce a novel approach that extends history-based tag suggestion with a visual content analysis crowd-sourced from YouTube: In our system, two similarity matchings are conducted for the questioned video, once with a large-scale reference dataset of user-tagged content and once with content in the user's personal history. The nearest neighbors found in both cases are merged based on their similarity, and a tag transfer is conducted by voting. This approach offers the following advantages:

- **Employing history content**: The approach uses not only the tags in a user's history but also the content. This can help disambiguate tag suggestion (think of multi-modal histories, i.e. users with multiple dominant interests).
- **Crowd-sourcing as a fall-back**: In cases where the user's history provided little or misleading information (e.g., in case of topical shifts) the large-scale content matching is implicitly used as a fall-back solution.

We present experiments on a dataset of 23,000 YouTube clips in which we demonstrate that the presented crowd-sourced *personalized tag transfer* gives strong improvements over a standard content-based one, and moderate improvements over a content-free history-based system.

2. RELATED WORK

In this section related work in the context of tag recommendation in general and tag recommendation of content-based systems is outlined.

Tag Recommendation.

General tag recommendation can be understood as inferring tags based on information sources such as text, global tagging behavior or user history. TagAssist [20] – a text-based tag recommendation system – employs TF-IDF to infer tags based on similar post from a corpus of known posts. More flexible systems are presented in [16], where tags provided by the user are extended by global co-occurrence derived from tagged images taken from Flickr. Regarding the utilization of user information, different approaches have been presented. In [15] a clustering-based approach on user interest is used for tag suggestion, whereas in [3] a random walk over ratings and tags of a user profile is performed. Extending simple user profiles, in [13] a user's social structure is utilized for personalized tag recommendation. History-based models have been introduced in [9, 14] where a user's history indicates a strong or even outperforming ability.

Content-based Tag Recommendation.

As our focus is particularly on personalizing *content-based* methods, we also review related work in this area. The automatic inference of tags from images and videos (referred to as *image/video annotation* or *concept detection*) has been study to extensive research for over a decade now (for a survey, please refer to [19]). A wide part of activities centers around collective quantitative efforts like TRECVID [18] or the PASCAL Visual Object Challenge [6], where evaluations are conducted on common datasets. Thereby, a widely used approach is to choose a *patch-based* image representation (for example, using *bag-of-visual-words* features [17] or their variants) and applying Support Vector Machines (SVMs) [19] for classification. As an alternative to SVMs, nearest neighbor methods have been investigated, which conduct a label transfer between the targeted content and a labeled training set using a similarity-based matching [12, 22], which offers the advantages of scalability to very large quantities of training data and transparency of the inference process. In this work, we choose a nearest neighbor setup, particularly because it allows an adaptation to user's personal content without an time-consuming re-training.

Less work can be found on adapting tag suggestions to user's contexts. Datta et al. [4] use a light-weight adaptation by a post-processing of concept scores to users' tagging behavior over time. In contrast to this work, our approach adapts to the *appearance* of a users personal content as well. From a broader perspective, our work is also related to methods embedding content-based inference with various types of context information (such as geo-tags [10], events [7], interest groups [24], or textual meta-data [21]).

3. APPROACH

To automatically suggest meaningful tags for a newly uploaded video, we describe multiple tag suggestion systems that rely on different modalities. Each such tag suggestion system takes a video v_{new} and associated information available via YouTube (e.g., the uploading user $u_{v_{new}}$). It suggests a list of tags together with a confidence score for each tag. For a video v its set of tags is denoted as T_v. The collection of all tags in the dataset (i.e., the union of all T_v) is denoted as $\mathcal{T}$.

3.1 Baselines

In the following we introduce several systems based on the following modalities:

History.

The history-based system utilizes tags that the user u used for previously uploaded videos – called the history H_u – to infer tags for a new video. For this the tags in H_u are sorted by their frequency (the most frequent one at the top) and this ordered list of tags is then suggested to the user. The score for each tag t is the frequency $H_u(t)$, normalized by the number of videos in the history (referred to in the following as the history's length). Figure 2 illustrates the distribution

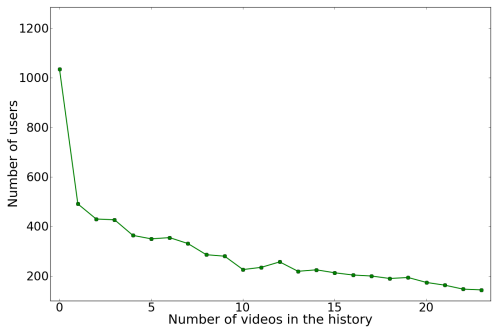

Figure 2: The frequency distribution of the number of videos in YouTube user histories. It follows a power-law distribution and shows that about 96% of the users have a non-empty history.

of the different history lengths over the users. There it is shown that although the history length follows a power-law distribution, only about 1,000 of the 23,000 users do not have any history at all. As about 96% of the users have a non-empty history this source of information is of sufficient density.

Co-Occurrence.

This system utilizes tags co-occurring with the user's history. This is motivated by the idea that having an explorative feature might be beneficial. For each tag t in the user's history all other videos on the platform (or a reasonable subset thereof) are tested if they are labeled with t. If so, all of the video's tags (except t) get a vote. The total number of votes for each co-occurring tag c can be calculated as $votes_c = \sum_{t \in H_u} H_u(t) \cdot cooc(t,c), \forall c \in \mathcal{T}$ where $cooc(t,c)$ denotes the number of times t and c occur together in other videos. To get the confidence scores for the tags, the votes are normalized by the number of videos considered, multiplied with the history length. This approach does not suggest any tags if the history is empty, as it is the case for the history-based system.

Channels.

Channels on YouTube are public collections of videos made available by an author. Other users may subscribe to channels and are then informed if new videos are uploaded to the channel. Channels allow to define a connection between $u_{v_{new}}$ and the authors A of his/her subscribed channels and therefore can be used as exemplary social signals. The votes for a tag t can be calculated as $votes_t = \sum_{a \in A} H_a(t)$, meaning that the occurrences of tags are summed up over the histories of all subscribed channels' authors. To get confidence scores the votes are divided by the sum of all authors' history lengths. If A is empty, i.e. the user has not subscribed to any channels, then no tags can be suggested. This system can be generalized to work with any other user-to-user(s)-connection (e.g., A is the set of friends).

Visual Content.

Two systems will be described, both rely on the same pipeline for label transfer, namely k Nearest Neighbor [5] using SIFT-based bag-of-visual-words features [11, 17]. Each feature represents a video keyframe extracted by a change detection algorithm [2]. For each keyframe there exist a set of k nearest neighbors drawn from the training set, which is independent of the test set. Two types of labels are used to suggest tags:

1. **Tag Transfer:** The *tags* of the nearest neighbor's associated video v_{nn} are used as labels, denoted as $T_{v_{nn}}$. $T_{v_{nn}}(t)$ denotes the number of times a tag t occurs in $T_{v_{nn}}$. For each keyframe *key* in the set of keyframes $K_{v_{new}}$ that represent v_{new}, its k nearest neighbors NN_{key} give votes for their tags. The votes are weighted by a weight w that depends on the nearest neighbor's rank in NN_{key}, denoted as $rank(nn)$. The rank for the most similar nearest neighbor is 1 and results in the highest weight. The number of votes t gets in total is:

$$votes_t = \sum_{nn \in NN_{key}} \underbrace{(|NN_{key}| - rank(nn))}_{w} \cdot T_{v_{nn}}(t)$$

   The votes are then summed up over all keyframes. To calculate the confidence score the votes are normalized by:

$$|K_{v_{new}}| \cdot \sum_{i=1}^{|NN_{key}|} i$$

2. **Vocabulary:** As an alternative to a direct tag transfer we study a predefined concept vocabulary, using the *concept* c of a nearest neighbor as its label. For this we have defined a set of 230 concepts like `baseball` or `shipwreck`. A nearest neighbor is labeled with concept c if its corresponding video was crawled from YouTube using c as query. To use this information for suggesting tags, a concept vocabulary is built for every concept c. For this the tags of all videos in the training set associated with c are summed up and ordered by frequency. This approach uses the training videos' labels to determine the concept $c_{v_{new}}$ of the newly uploaded video. v_{new} is then suggested the concept vocabulary for the concept $c_{v_{new}}$. The confidence scores can be calculated by normalizing the concept's tags by the number of videos that are labeled with this concept.

3.2 Visual Personalized Tag Transfer

Our key contribution is the *Visual Personalized Tag Transfer* approach which seeks to combine both history-based signals as well as content-based signals. This might be especially beneficial if the history of a user is not coherent, i.e. tags cover more than one topic. In such a case the content-based signals can help to boost tags that are about the topic present in the uploaded video. Furthermore, this enables the system to suggest tags even if the user has no history.

The working mechanism of this approach for a single keyframe $key \in K_{v_{new}}$ can be seen in Figure 3. There it is shown that the videos in the user's history are represented by keyframes as it is the case for the content-based systems. The history keyframes are then re-ranked according to their similarity to *key* (again calculated using Bag of Visual Words with SIFT features). As the nearest neighbors gained by the Tag Transfer approach (in the following re-

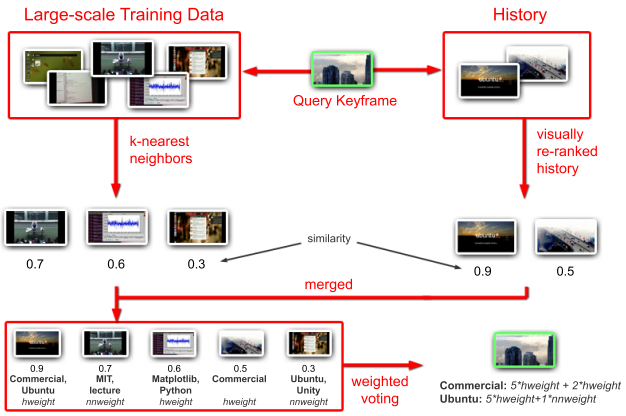

Figure 3: The Visual Personalized Tag Transfer approach to automatic tag suggestion. For this approach the history is merged with global nearest neighbors.

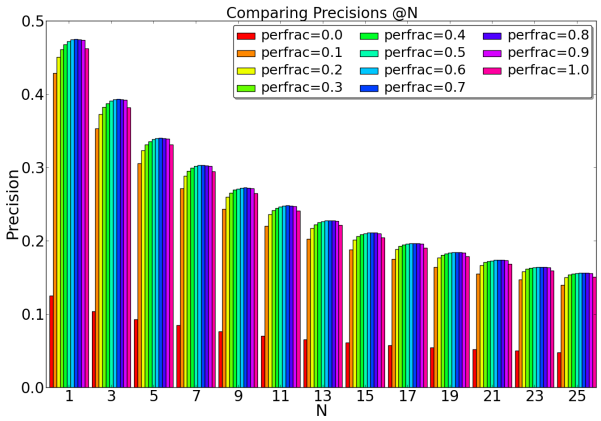

Figure 4: The Precision of the Visual Personalized Tag Transfer system for different values of *perfrac*.

ferred to as the *global* nearest neighbors) use the same similarity measure, both the re-ranked history RH_{key} and the global nearest neighbors can be merged and again be ordered by their similarity. This merged list is denoted as M in the following. To adjust the influence of both subsets a parameter called the *personalization fraction*, denoted as *perfrac*, is introduced. From this parameter two weights are calculated as the minimal integer solution of $perfrac = \frac{|H| \cdot hweight}{|NN_{key}| \cdot nnweight}$. Each entry in M gives a number of votes that is the product of the inverse of its rank and *nnweight* for the global nearest neighbors or *hweight* for tags of the history keyframes respectively. This can be calculated by:

$$votes_t = hweight \cdot \sum_{nn \in M \cap RH_{key}} rank(nn) + nnweight \cdot \sum_{nn \in M \cap NN_{key}} rank(nn)$$

where $rank(nn)$ denotes the rank of nn in M. For the final result the votes for all tags are summed up over all keyframes $key \in K_{v_{new}}$.

4. EXPERIMENTS

In the following the presented systems will be evaluated quantitatively and compared to each other. For this we will describe the test setup in Section 4.1, as well as the underlying concept detection pipeline of the content-based systems. After this, in Section 4.2 the accuracy of all systems will be compared and the performance of the concept detection will be shown and discussed. Preliminary work and additional experiments can be found in [8].

4.1 Setup

We selected a vocabulary of 230 concepts. For evaluation purposes 200 videos were crawled for each of the 230 concepts like `baseball` or `shipwreck` (for more details see [23], which used the dataset for video-retrieval using visual and and semantic signals). For each concept 100 videos have been randomly chosen for training purposes and 100 for the actual testing, both sets sharing no common entries. This results in 23,000 test videos, on which both the concept detection pipeline and the tag suggestion system are evaluated.

The suggested tags are evaluated against the tags - neglecting stop words - that were assigned to the test videos by the respective original uploader on YouTube. As performance measures we use Precision and Recall of the suggested tags in terms of the tags the video was labeled with. The motivation behind this is that user studies, which would be an alternative, are easily deluded by general terms [21] and may not catch the original uploader's intentions (for example in [21] the average relevance of tags by the original uploader was found to be 0.43 when rated by third-party raters). The single Precision and Recall values are averaged over all 23,000 videos and reported over different ranks N, i.e. 1...25.

For the concept detection pipeline roughly 3 million keyframes were extracted. To save computational time we subsampled the keyframes randomly to 1,000 per concept (at least one for each video). For each of these keyframes (scaled to 250 by 250 pixels) a descriptor was build using a codebook of 3,000 visual words. Further, for the nearest neighbors matching k was set to $k = 100$ motivated by the trade-off between computational time and choosing the square root of the size of the dataset as a good value for k [5].

4.2 Results

We first focus on the Visual Personalized Tag Transfer approach: as its performance depends of the personalization fraction, different values for *perfrac* can be seen in Figure 4. It shows that values between 0.6 and 0.8 work best, especially 0.7 works well. This means that we have a quite broad range for *perfrac* in which the system performs well, making the choosing of the parameter less difficult. It also shows that both extreme cases 0.0 (relying on the global nearest neighbors only) and 1.0 (relying on the history only) work worse than these values.

In Figure 7 the performance of all content-based systems is illustrated. It can be seen that **TagTransfer** performs better than **Vocabulary(real)**, which suggests concept names as tags. This is due to the limited accuracy of the concept detection pipeline which achieves a Mean Average Precision of only about 5.33% (random guessing would result in 0.4%). This is also shown by the performance of an oracle test run **Vocabulary(oracle)**, using a perfect concept detection pipeline illustrating that users use at least some

similar words for the same concept. Figure 8 shows the performance of all presented systems in terms of Precision. From these we can see following results:

- The **History**-based system proves to be a strong approach and is able to outperform most other systems. This confirms earlier findings [14] and illustrates that that the users in the dataset tend to have consistent histories.
- The YouTube Data API[3] does not allow us to distinguish if a users has no channels or has just set this information to private. Because of this, the **Channel**-based system is evaluated twice: Once on all users (*all*) assuming a Precision and Recall of 0 for users with inaccessible channels (true for 44% of all users) and once only on those users who have channels (*wc*). The evaluation of (*all*) resulted in the worst performance of all systems. But under the assumption of (*wc*) tag suggestion performs much better, being among the best single-modal system for lower ranks of N. This shows that if channels are available they can reflect the user's interests. Nevertheless both systems perform worse than the history-based system.
- The **Co-Occurrence**-based system is able to outperform all other realistic single-modal systems for ranks N greater 1, except for the history-based system which performs better. This illustrates that tags which occur together with the ones from the user's history at least partly fit the user's content.
- The best purely **Visual Content**-based system – the Tag Transfer system – works worse than most other systems. This shows that the global nearest neighbors alone are not reliable enough for good tag suggestion.
- The **Visual Personalized Tag Transfer** system with a personalization fraction of 0.7 greatly outperforms all other non-oracle content-based systems. It also modestly outperforms the history-based system, again showing that enhancing the user's history with content-based signals can in fact help to improve the quality of the tag suggestion.

The same results can be observed when considering Recall rather than Precision. The benefits of the Visual Personalized Tag Transfer over a purely content-based system, in this case the Tag Transfer approach, can be seen in Figure 5: although tags like `fireworks` and `paris` can be inferred correctly when using visual information alone, such a system is also deluded by the flickering lights of a TV show resulting in tags like `show`. The personalization of these results by the user's history, as done by the Visual Personalized Tag Transfer system, helps to properly suggest tags being present visually - a perfect result even though the history is inconsistent (including videos about fireworks *and* cars). The impact of this inconsistence is damped by the fact that the history is re-ranked accord to visual similarity. Another advantage is the detection of interest change: for example a user who only tagged his past videos with `home video` but now uploads a video about snooker is only suggested `home video` by the history-based system, whereas the Visual Personalized Tag Transfer system correctly suggests `snooker`.

[3] http://gdata.youtube.com/demo/index.html

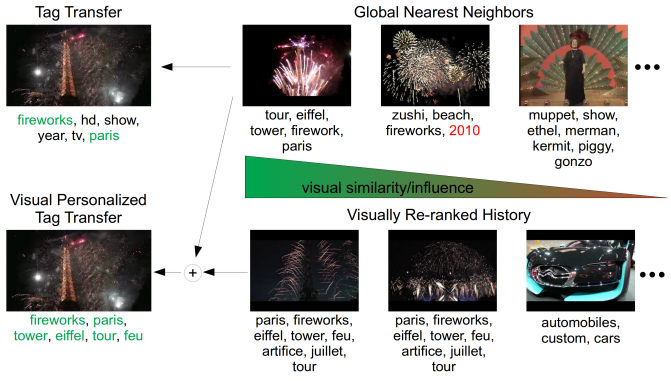

Figure 5: An example for the different performances of the Tag Transfer approach (top) and the Visual Personalized Tag Transfer approach (bottom). Green tags are suggested correctly, red ones are stop words. Each video is represented by a single keyframe and its tags.

Examples for actual suggestions of the systems can be seen in Figure 6. In the first row are videos illustrated by a keyframe, in the second row the tags by the uploading user are shown and in the bottom row the top six suggested tags of the respective system can be seen. Green tags are suggested correctly, red ones are stop words (not considered for the performances measures) and orange ones are tags that fit but are not in the original tags.

It should be considered that we have chosen to evaluate the systems on the tags assigned by the original user to get a realistic measure for the usefulness of the suggested tags. This might penalize more exploratory systems, although such a property might be desirable, as such systems are more likely to suggest fitting tags the user did not think of. Some examples for such cases can be seen in Figure 6: the Tag Transfer system suggests `game` for a video about a video game that was not tagged with this word.

5. CONCLUSION

In this paper we tackled the challenge of automatic tag recommendation for web video clips. First, we evaluated in a quantitative experiment on 23,000 YouTube videos different tagging strategies such as user upload history-based ones and social-based ones. Further, we introduced a novel approach, which fuses personalized tag suggestion with crowsourced content matching. This way, content analysis can help to improve tag suggestions indicating strong improvements over a standard content matching, and moderate ones over a content-free history-based ranking.[4]

6. REFERENCES

[1] M. Ames and M. Naaman. Why we Tag: Motivations for Annotation in Mobile and Online Media. In *Proc. SIGCHI HFCS*, 2007.
[2] R. Brunelli, O. Mich, and C.M. Modena. A Survey on the Automatic Indexing of Video Data. *J. Vis. Com. and Im. Rep.*, 10(2), 1999.
[3] M. Clements, A.P. De Vries, and M.J.T. Reinders. The task-dependent Effect of Tags and Ratings on Social Media Access. *ACM TOIS*, 28(4), 2010.
[4] R. Datta, D. Joshi, J. Li, and J. Wang. Tagging over Time: Real-world Image Annotation by Lightweight Meta-Learning. In *Proc. MM*, pages 393–402, 2007.

[4] This work was funded by the Google Research Awards Program.

	History	Co-Occurrence	Channel	TagTransfer	Personalized TagTransfer
					
Original User's Tags	mount, everest, maldives, himalayan, dreams	high, stakes, poker, daniel, negreanu, vs, dwan	dance, salsa, fiesta, 2007, oliver, and, luda	simpsons, hit, run, mission, petty, theft, homer	michael, jordan, nike, gatorade, commercial, air
Suggested Tags	himalaya, nepal, tibet, himalayan, dreams, hollywood	high, hd, poker, definition, stakes, school	salsa, dance, mambo, congress, dancing, dvd	hit, simpsons, run, homer, race, game	michael, jordan, air, commercial, nike, game

Figure 6: Examples of suggested tags for the described systems. For each video (illustrated by a keyframe) the original user's tags and the suggested tags are shown. Red tags are stop words, green ones were suggested correctly and orange ones fit the video but were not used originally.

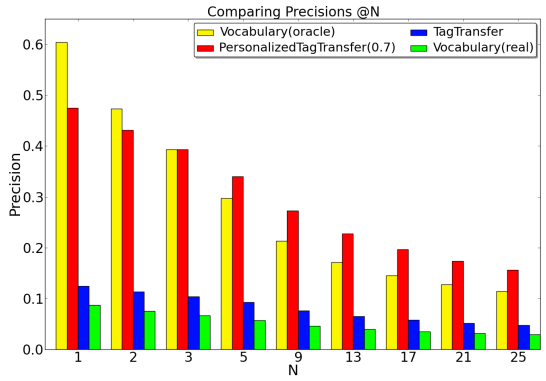

Figure 7: Comparing the Precisions of all described content-based tag suggestion systems.

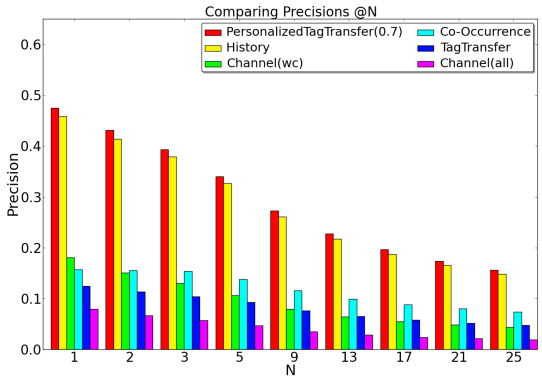

Figure 8: Comparing the Precisions of all presented tag suggestion systems (the purely content-based approaches are represented by the Tag Transfer approach only).

[5] O. Duda, R, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 88(2):303–338, 2010.

[7] A. Gallagher, C. Neustaedter, L. Cao, J. Luo, and T. Chen. Image Annotation using Personal Calendars as Context. In *Proc. ACM Multimedia*, 2008.

[8] D. Henter. Combining Social and Content Based Signals for Personalized Tag Suggestion on YouTube. Bachelor's thesis, University of Kaiserslautern, 2012. https://madm.dfki.de/_media/2012_bachelors_thesis_dominik_henter.pdf.

[9] X. Li, E. Gavves, C. G.M. Snoek, M. Worring, and A. W.M. Smeulders. Personalizing Automated Image Annotation using Cross-Entropy. In *Proc. ACM Multimedia*, 2011.

[10] X. Li, C. Snoek, M. Worring, and A. Smeulders. Fusing Concept Detection and Geo Context for Visual Search. In *Proc. ICMR*, 2012.

[11] D. G. Lowe. Object Recognition from Local Scale-Invariant Features. In *Proc. ICCV*, 1999.

[12] A. Makadia, V. Pavlovic, and S. Kumar. A New Baseline for Image Annotation. In *Proc. ECCV*, 2008.

[13] A. Rae, B. Sigurbjörnsson, and R. van Zwol. Proc. RIAO. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, 2010.

[14] N. Sawant, R. Datta, J. Li, and J. Z. Wang. Quest for Relevant Tags using Local Interaction Networks and Visual Content. In *Proc. MIR*, 2010.

[15] A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke. Personalized Recommendation in Social Tagging Systems using Hierarchical Clustering. In *Proc. ACMRS*, 2008.

[16] B. Sigurbjörnsson and R. van Zwol. Flickr Tag Recommendation based on Collective Knowledge. In *Proc. WWW*, 2008.

[17] J. Sivic and A. Zisserman. Video Google: Efficient Visual Search of Videos. In *Toward Category-Level Object Recognition*, pages 127–144, 2006.

[18] A. Smeaton. Large Scale Evaluations of Multimedia Information Retrieval: The TRECVid Experience. In *Proc. CIVR*, pages 11–17, 2005.

[19] C. Snoek and M. Worring. Concept-based Video Retrieval. *Foundations and Trends in Information Retrieval*, 4(2):215–322, 2009.

[20] S. Sood, S.. Owsley, K. J. Hammond, and L. Birnbaum. TagAssist: Automatic Tag Suggestion for Blog Posts. In *Int. ICWSM*, 2007.

[21] G. Toderici, H. Aradhye, M. Pasca, L. Sbaiz, and J. Yagnik. Finding Meaning on YouTube: Tag Recommendation and Category Discovery. In *Proc. CVPR*, 2010.

[22] A. Torralba, R. Fergus, and Y. Weiss. Small Codes and Large Databases for Recognition. In *Proc. CVPR*, pages 1–8, 2008.

[23] A. Ulges, M. Koch, D. Borth, and T. Breuel. TubeTagger – YouTube-based Concept Detection. In *Proc. Int. Workshop on Internet Multimedia Mining*, December 2009.

[24] A. Ulges, M. Worring, and T. Breuel. Learning Visual Contexts for Image Annotation from Flickr Groups. *IEEE Transactions on Multimedia*, 13(2):330–341, 2011.

[25] YouTube Blog: YouTube Videos now served in WebM. available from http://youtube-global.blogspot.com/2011/04/mmm-mmm-good-youtube-videos-now-served.html (retrieved: Sep'11), 2011.