

Uploader Intent for Online Video: Typology, Inference, and Applications

Christoph Kofler, Subhabrata Bhattacharya, *Member, IEEE*, Martha Larson, *Member, IEEE*, Tao Chen, Alan Hanjalic, *Senior Member, IEEE*, and Shih-Fu Chang, *Fellow, IEEE*

Abstract—We investigate automatic inference of *uploader intent* for online video, i.e., prediction of the reason for which a user has uploaded a particular video to the Internet. Users upload video for specific reasons, but rarely state these reasons explicitly in the video metadata. Information about the reasons motivating uploaders has the potential ultimately to benefit a wide range of application areas, including video production, video-based advertising, and video search. In this paper, we apply a combination of social-Web mining and crowdsourcing to arrive at a typology that characterizes the uploader intent of a broad range of videos. We then use a set of multimodal features, including visual semantic features, found to be indicative of uploader intent in order to classify videos automatically into uploader intent classes. We evaluate our approach on a dataset containing ca. 3K crowdsourcing-annotated videos and demonstrate its usefulness in prediction tasks relevant to common application areas.

Index Terms—Crowdsourcing, indexing, search intent, video audience, video popularity, video search, video uploader intent.

I. INTRODUCTION

THE challenge addressed in this paper is to infer, automatically, on the basis of multimodal features, the reasons motivating users to upload videos to the Internet. Specifically, we focus on classes of *uploader intent* that capture the major reasons why users create videos and post them online. The concept of uploader intent is illustrated by two pairs of videos in Fig. 1, exemplifying types of uploader intent. Pair (a) are videos that communicate birthday greetings and engagement congratulations, and serve as examples of videos uploaded in order to convey emotion, i.e., *express affect*. Pair (b) are videos that provide instructions on how to program a transmitter and to repair a motorcycle, and serve as examples of videos uploaded to *explain*. The contrast between the textual description of the two videos in each of the two pairs demonstrates that in some cases users state their intent in uploading a video in the metadata (first

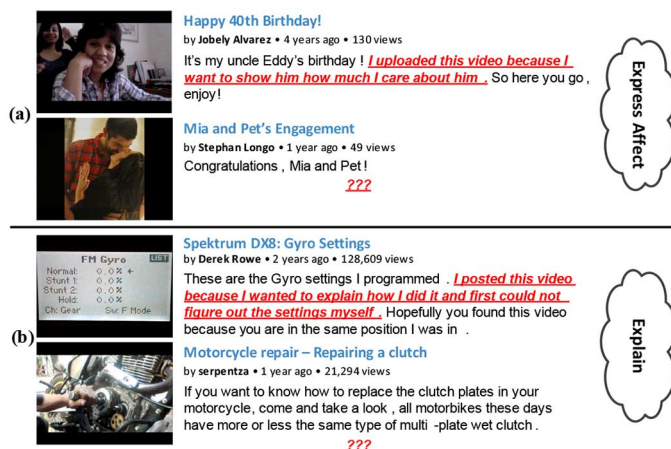


Fig. 1. Two pairs of videos uploaded to the Internet for different reasons, either (a) to *express affect* or (b) to *explain*. Only the first video in each pair contains an *explicit* statement in its description why a user uploaded it ('I uploaded/posted this video because ...') [Source: YouTube].

video in each pair), while in others they do not (second video in each pair). The fact that the vast majority of videos are uploaded by users who do *not* explicitly express their intent means that automatic inference of user intent is necessary if this information is to be exploited in application scenarios.

Our investigation of uploader intent for online video is motivated by the wide variety of application areas that ultimately stand to benefit from information on the reasons which prompt users to upload videos. These areas cover a diverse spectrum including video production and video search. For example, in the area of video production, knowledge about uploader intent could improve video authoring tools, guiding the user in producing videos with a fitting 'look and feel', for instance, by automatically recommending editing templates or Instagram-like filters. Another example is that uploader intent could aid the automatic matching of advertisements to videos, by providing information concerning the intended target audience of a video.

In this paper, we explore uploader intent in the context of two prediction tasks related to potential applications in the area of video search: we use inferred uploader intent to predict the size of the audience that the uploader aimed to reach with the video (designated *reach*), and to infer statistics related to the popularity of the video (designated *impact*). Both of these tasks can also contribute to refining video search result lists, e.g., by reranking to move videos most likely to satisfy the user into top ranks. Further, in the context of video search, matching the intent that motivates user queries, i.e., the *search intent* [1], [2] (e.g., 'how to build a house'), with the intent that motivates a

Manuscript received September 24, 2014; revised March 20, 2015; accepted May 30, 2015. Date of publication June 15, 2015; date of current version July 15, 2015. This work was supported by the Dutch national program *COMMIT*. This work was carried out when C. Kofler was visiting Columbia University. The work of C. Kofler was supported by the Google Europe Doctoral Fellowship in Video Search. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Xiao-Ping Zhang.

C. Kofler, M. Larson, and A. Hanjalic are with the Delft University of Technology, Delft 2628CD, The Netherlands (e-mail: c.kofler@tudelft.nl).

S. Bhattacharya is with Imaging and Computer Vision, Siemens Corporation, Corporate Research, Princeton, NJ 08540 USA.

T. Chen and S.-F. Chang are with Columbia University, New York, NY 10027 USA.

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TMM.2015.2445573

user to upload a video (e.g., ‘explaining how to build a house’), would, a priori, appear to hold great promise for improving video search results lists. For this reason, our paper includes an additional small-scale user study on video search that explores whether or not the assumption that search intent correlates with uploader intent is justified and can be used in intent-aware video search results list optimizations [2].

The problem of automatic inference of uploader intent addressed in this paper is a challenging one. The nature of this challenge is highlighted by the examples in Fig. 1. It can be seen, that uploader intent spans videos that cover diverse topics. For this reason, topic detection approaches will *not* alone address the problem of uploader intent prediction. Further, videos expressing positive or negative emotion both fall into the category of videos that *express affect*. For this reason, the challenge of predicting uploader intent does *not* reduce to conventional affect detection. Given that most previous work in the field of video content analysis has targeted either topic or affect, analyzing video for uploader intent calls for innovative solutions outside mainstream video processing research. Our approach to inferring uploader intent is particularly useful because it does not require users to interact with the videos (i.e., producing a viewcount) or contribute additional information (e.g., comments, likes), nor does it require any information about the background of the uploading user (e.g., overall upload history, status as a private user or a public company). Consequently, our approach can be applied immediately during the upload process in a computationally inexpensive manner.

The purpose of this paper is to lay the groundwork necessary for effective deployment of the concept of uploader intent for online video, and to demonstrate its relevance to potential applications. We conduct a thorough study of uploader intent (Section III), which includes a social-Web mining procedure and a crowdsourcing user study. The result is a typology of important classes of uploader intent for online video. We then address the challenge of uploader intent inference (Section IV and V). The novelty of our classification approach lies in two intent-specific aspects. First, the approach learns uploader intent from videos on the basis of multimodal (both visual and textual) features determined to be indicative of uploader intent. Second, it applies to general videos on Internet video sharing platforms such as YouTube, which we take as representative. Finally, we turn to application areas (Section VI), investigating the tasks of predicting audience size, i.e., *reach*, and popularity, i.e., *impact*, and also the link between uploader intent and *search*. Section VII concludes and provides an outlook on future work.

The contributions of the paper are experimental results answering the following research questions:

- *RQ1*: What are main uploader intent classes constituting a typology useful to cover a wide range of videos?
- *RQ2*: Are features extracted from textual and visual data associated with videos indicative of uploader intent?
- *RQ3*: Can the class of uploader intent of a video be predicted automatically?
- *RQ4*: Can the size of the audience at which the uploader targeted the video be predicted automatically using uploader intent? (*reach*)
- *RQ5*: Can popularity characteristics of videos be predicted automatically using uploader intent? (*impact*)

- *RQ6*: Does a correlation plausibly exist between video uploader intent and query search intent? (*search*)

II. RELATED WORK

A. User Intent

Several studies have investigated why users create, share and upload multimedia content. Kindberg *et al.* [3] study, by means of interviews, why users take pictures on mobile devices. They find that pictures were taken for *sharing* and *personal* use, and for *affective* and *functional* use. Lux and Huber [4] perform an exploratory study using semi-structured interviews to investigate the types of user intent underlying video production. They arrive at the following set of intent categories: *preservation*, *sharing*, *affection*, *functional* and *technical interest*. Campanella and Hoonhout [5] investigate why users capture home videos and find that main reasons are to *keep memories of someone’s life* and to *share experiences* with family members and friends. Borneo and Barkhuus [6] study motivations for video microblogging, and discover that the main goals of bloggers are *self-expression*, *entertainment* and *self-presentation*. Park *et al.* [7] conduct surveys to investigate factors that are associated with users’ intentions in uploading videos to the Internet. Their results reveal that, in particular, ego-involvement (e.g., self-presentation) is associated with users’ attitudes toward uploading behavior.

In this paper, we study uploader intent for video. We find similarities with previous work, suggesting that there is an element of sharing behavior that is fundamental, perhaps related to human nature. Our work achieves insight that is up to date—important due to the huge changes that have occurred in the sophistication and availability of capture devices to a broad population of users. Crucially, these insights are also specific to video, as will be discussed in detail in the paper. We apply a social-Web mining approach to discover classes of uploader intent. The approach follows a methodology similar to one that has proven effective in earlier work on the discovery of *search intent* classes for video search [1]. Our approach is superior to conventional methods such as those using transaction log analysis [8], interviews [4] or surveys [7] because it exploits evidence from a very large user population to directly access spontaneously expressed information about why users upload videos. In a further step, we use a crowdsourcing user study, which gives us access to a large number of users, both to refine initial classes discovered through social-Web mining and to annotate videos based on our uploader intent typology.

B. Video Understanding

Analysis and understanding of videos has been approached from different perspectives, such as what a video depicts, what it is about, where it was taken, and what types of emotions it elicits in users [9]. A large amount of research effort has been invested in automatically inferring visual concepts from images and videos (cf. [10] etc.). This line of research is critically necessary in order to better understand the content of videos, what they depict, but also what topics they cover. For example, Rudinac *et al.* [11] represent long videos by the distribution of visual

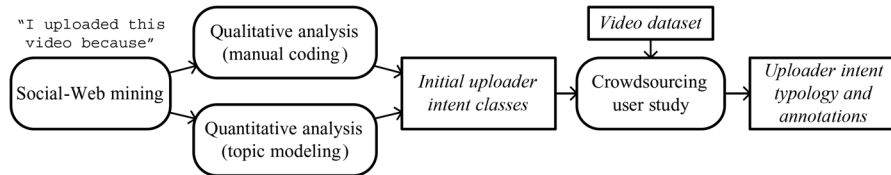


Fig. 2. Our procedure for establishing a video uploader intent typology. A social-Web-based mining approach delivers reasons why users upload videos to the Internet, which are subjected to both qualitative and quantitative analysis in order to identify initial intent classes. These classes are then refined and extended by a crowdsourcing study to arrive at a final set of uploader intent classes and an annotated video dataset.

concepts appearing in them to infer information about their general subject matter. Schmiedeke *et al.* [12] automatically infer the user-chosen topic category of consumer-produced videos based on their textual and visual features. Chen *et al.* [13] automatically discover semantic concepts that are related to complex events in videos. Borth *et al.* [14] automatically infer semantic concepts that are associated with emotions conveyed in images. Related to the task of automatically tagging videos [9], recent work such as that of Habibian and Snoek [15] automatically generates human-readable descriptions based on the analysis of the visual channel of videos.

Our work is similar to these approaches in that it adds to the understanding of videos, and also aims to infer additional information from them. Also similar is that it applies a standard classification approach to perform the prediction step for unseen videos. The critical difference is that our work explicitly analyzes videos from the perspective of why users originally uploaded them to online video sharing platforms. We apply textual features associated with videos, as well as low-level features and mid-level semantic representations extracted from videos that we found indicative of uploader intent in order to classify videos into the intent classes that we discover.

C. Online Video

A number of studies have been carried out in order to analyze characteristics associated with videos once they are available on online video sharing platforms. These investigations focus on understanding and predicting the popularity a video will achieve among viewers and also their commenting behavior, as well as examining which users are interested in which type of videos. For example, Siersdorfer *et al.* [16] investigate and automatically predict commenting and comment rating behavior of users on YouTube. Cha *et al.* [17] examine the distribution and evolution of the popularity of YouTube videos, as well as related user behavior. Figueiredo *et al.* [18] analyze how the popularity of individual YouTube videos evolves since the moment they are uploaded. Similar to Szabo and Huberman [19], Pinto *et al.* [20] automatically predict a video's future popularity based on early popularity measures. Cheng *et al.* [21] perform an analysis of characteristics of YouTube videos and their evolving social network including length of videos belonging to different categories and growth trends in uploading, views and ratings.

In our work, we exploit the inferred uploader intent of videos to automatically predict the size of their intended target audience (*reach*), as well as the activity they generate in terms of viewcount, likecount and ratings (*impact*). Work on prediction of audience characteristics is Weber *et al.* [22] and Abisheva *et al.* [23], who carry out a text-based analysis of Twitter and YouTube data to study the relationships between the timing and

topical patterns of video sharing and the demographic characteristics of users, including political viewpoints. We apply similar features as baselines to evaluate our popularity and target audience prediction approaches. However, the difference is that our work goes beyond only text features. Further, it focuses on the aim of the uploading user to reach a specific audience (e.g., a few friends vs. the Internet at large) with a particular video, rather than on specific characteristics of that user. Instead of investigating the intended audience as we do, they investigate who watches which type of content on YouTube, and automatically predict the partisanship of political YouTube videos based on their textual metadata.

III. VIDEO UPLOADER INTENT TYPOLOGY

Here, we address question *RQ1*, ‘What are main uploader intent classes constituting a typology useful to cover a wide range of videos?’. To ensure that our typology is sufficiently general, i.e., applicable to the general case of videos on Internet sharing platforms, we adhere to a specific set of criteria. These criteria have previously been successfully applied to define a useful typology for search intent classes [1]. The criteria are: the typology should be *complete* (i.e., it should provide good coverage of a wide range of video on the Internet), *intuitive* (i.e., classes should be clear, recognizable and distinctive), and *compact* (i.e., the classes should be comparable in their importance/coverage). Our approach to typology definition is illustrated in Fig. 2. This section explains the process in detail.

A. Social-Web Mining for Video Uploader Intent Discovery

In order to find the initial set of uploader intent classes, we follow a data-driven approach. We define a regular expression capturing a common word pattern used by uploaders to express intent in the metadata of their videos: “I {uploaded | shared | posted} {this | the} video because” We use strings generated by this regular expression to query YouTube and Bing video search. The result was a collection of descriptions associated with 4,351 unique videos.

We carry out a qualitative analysis via a manual coding process, as also applied in, e.g., [1], [24], to a subset of the data that is small enough to handle by hand. The process allows us to discover classes in the data in a bottom-up fashion. We individually inspect a randomly selected subset of 150 reasons in our collection of descriptions. For each reason, we decide whether it fits a previously discovered initial uploader intent class or if it represents a new, yet unseen category. We repeat this process by comparing and contrasting discovered classes, merging and dividing the clusters until each contains descriptions that characterize a single type of uploader intent, which is different from that represented by the other clusters.

We extend this manual analysis with an automatic analysis that allows us to cover the complete set of data. We apply Latent Dirichlet Allocation (LDA) [25] to our set of 4,351 video descriptions, and compare the resulting clusters with the clusters resulting from the manual coding of the 150 videos. Our goal is to verify the initially discovered classes, and discover new classes that the initial manual analysis may have missed.

The two steps resulted in an initial set of uploader intent classes: *Request from acquaintance*, *First upload of content*, *Better quality or preference of material compared to existing content*, *Explanation or illustration*, *Clarification for community*, *Asking for an opinion*, *Expressing an opinion*, *Expressing affect of entertaining*, *(Self-)promotion or promoting an article*, *Continuous uploader or seeking for subscribers*, *Random/Bored*, and *Backing up or testing*.

Examination of these classes reveals that they share high-level similarity with the typology developed by Kindberg *et al.* [3], who carried out interviews to determine why people take pictures with mobile devices. As Kindberg *et al.* [3], we observe that some of our initial classes are related to *functional* motivations (e.g., *Explanation or illustration*), while others focus on *affective* motivations (e.g., *Expressing affect or entertainment*). In addition, some of the classes contain videos which are uploaded to be *shared with others* (e.g., *Expressing an opinion*) or for *individual use* (e.g., *Backing up or testing*). These classes build the foundations for our target audience classes defined and used later in this paper.

In the next section, we describe the crowdsourcing user study with which we make the next step towards the final uploader intent typology.

B. Crowdsourcing for Video Uploader Intent Discovery

The crowdsourcing user study refines the initially discovered classes by integrating the perspective of a large number of Internet users. The study ensures that our typology is intuitive, in other words, that our uploader intent classes are generally recognizable as applicable to videos on the Internet, and that it covers a wide range of online video. The study is carried out on a manually selected set of 40 videos that do not include explicit statements of uploader intent in their descriptions, and are spread so as to reflect diversity of video on YouTube. These videos are sampled from a larger video dataset that we describe in further detail in Section III-C.

Our crowdsourcing study was implemented as a Human Intelligence Task (HIT) on Amazon Mechanical Turk (AMT). Each video was presented to workers in an individual HIT, along with the initial set of uploader intent classes. Workers were asked to browse each video and read the associated metadata, and then choose the class that they found to be the best fit. We asked them to make the decision by projecting themselves into the position of the original video uploader. We explicitly asked the workers to focus on *why* the person would upload videos to YouTube, independently of *what* the video is about or the emotional impact of the video. The workers are provided with a text box, and asked to suggest a new category in case the video did not fit into one of the classes in the initial set of uploader intent classes. After choosing the class, workers were asked to explain using 2-4 sentences their choice. These explanations provide us with insight into the workers' confidence in their choice, and also

serve as quality control, since the form and syntax of the answer reflects the seriousness with which the worker carried out the task. Each video was annotated by six workers.

Manual analysis of the worker responses yielded insight into frequently occurring reasons for which people upload videos. Some of our initial, rather specific classes, in particular *Request from acquaintance*, *First upload of content*, *Better quality or preference of material compared to existing content*, *Clarification for community*, *Continuous uploader or seeking for subscribers* and *Random/Bored* were not observed by the crowdsourcing workers. This observation suggested either that they are not highly frequent, and/or that they are not intuitive, and should not be included in the final typology. For the other classes, the workers' explanations were the source of valuable suggestions. The classes *Explanation or illustration* and *Expressing affect or entertainment* appeared to be too general, and cause confusion. We addressed this issue by refactoring our classes. This process resulted in a set of *Explaining* classes for videos that *convey knowledge* (i.e., provide declarative knowledge), *teach practice* (i.e., provide performative knowledge), *illustrate*, and also a set of *Communicating* classes including *express affect or interest*, but not including *Entertaining* videos. Additionally, workers observed that a relatively large number of videos were uploaded to share particular experiences. We accommodated this response by adding *Sharing* as a separate uploader intent class.

C. Dataset for Video Uploader Intent

Next, we turn to a description of how our video dataset was created. Note that this is the dataset that will be used for later classification experiments investigating automatic inference of intent. Our experiments call for a dataset fulfilling two requirements: it should provide good coverage of the diverse types of videos that people search for on the Internet, and it should also be annotated with uploader intent. Note that existing datasets such as [26], are not appropriate since they are created to study visual information in video, and our approach requires data that is not biased towards a particular modality. Collecting random videos from the Internet is also likely to be a suboptimal solution, as there is no steering towards covering an acceptable scope of uploader intents. To satisfy our requirements, we make use of an already existing, publicly available set of 935 queries¹ associated with diverse video-related information needs expressed by real-world users on a question answering forum [1]. We collect videos related to these queries, and then use a crowdsourcing experiment to collect uploader-intent labels for this dataset.

For the purpose of video collection, we again choose YouTube, as representative of an online video sharing platform and a state-of-the-art video search engine. To ensure the relationship of the videos to user information needs, we filter for 'unanswerable' needs by discarding queries that return less than 10 results, as in [2], leading to a set of 715 queries. Next, we sample five videos from the top-25 results returned from each query. We avoid choosing only the top-5 in order to preserve diversity with respect to popularity and quality aspects. We avoid videos below rank 25, in order to remain in the

¹[Online]. Available: <http://goo.gl/2NW6LP>

TABLE I
VIDEO UPLOADER INTENT CLASSES DISCOVERED BY OUR MINING AND CROWDSOURCING APPROACH (VERTICAL) AND THEIR MAPPING TO THE CLASSES OF OUR VIDEO UPLOADER INTENT TYPOLOGY (HORIZONTAL AND BOLD) USED FOR EXPERIMENTATION AND ANALYSIS IN THIS PAPER

Uploader intent classes resulting from crowdsourcing	Description Users upload the video to...
<i>Explaining (UI_{EX})</i>	
<i>Convey knowledge</i>	convey knowledge and to teach viewers of the video something they would also be able to learn when, for example, reading about it in a book or attending a lecture.
<i>Explain something practical</i>	explain viewers of the video something practically or to help them acquire a new skill.
<i>Illustrate</i>	illustrate viewers of the video something (e.g., to make something better understandable, to clarify something, or to (visually) present something).
<i>Sharing (UI_{SH})</i>	
<i>Share an experience or event</i>	share a (real-life) experience or event to viewers of the video (who, for example, might have missed this experience or who would like to relive this very particular experience).
<i>Promoting (UI_{PR})</i>	
<i>Promote</i>	promote something to viewers of the video (e.g., self-promotion for the uploader, sharing talent, promoting an article, bragging, to reach an audience, etc.).
<i>Communicating (UI_{CO})</i>	
<i>Express affect or interest</i>	express affect or an interest about the video's topic to its viewers (e.g., to express that the uploader is a fan of the video's content, etc.).
<i>Express an opinion</i>	express and share an opinion or idea with viewers of the video.
<i>Ask for an opinion/feedback</i>	ask for feedback/the opinion of the viewers of the video (e.g., about a particular skill, statement, action, opinion etc. portrait in the video).
<i>Entertaining (UI_{EN})</i>	
<i>Entertain</i>	purely entertain its viewers.

result range that can be assumed to be topically focused with state-of-the-art video search engines [27]. The result is a video dataset containing 3,575 videos, which has good diversity, and is also of a size tractable to annotate via crowdsourcing. For each video, we collect the metadata, including title, tags, description, genre categories, ratings, likecounts, viewcounts, and download three keyframes that are automatically extracted by YouTube.

D. Finalizing the Typology

As a result of our social-Web data mining step (Section III-A), whose output was refined by our crowdsourcing study (Section III-B), we arrive at a typology of nine uploader intent classes. The columns of Table I contain the labels and explanations for these classes.

In this section, we carry out a second, large-scale crowdsourcing study in which we asked crowdworkers to annotate the videos in the dataset described in Section III-C with these nine classes. We then carry out an analysis of these patterns of assignment in order to validate the typology. Validation involves checking that the typology fulfills the criteria set out at the beginning of this section, i.e., whether it is complete, intuitive and compact. On the basis of this analysis, we group the nine classes into the final uploader intent typology, constituting the five classes (presented horizontally and in bold) in Table I.

Our second, large-scale crowdsourcing study is again carried out on AMT. The HIT presents the crowdworkers with the 3,535 videos in the dataset (the original 3,575 minus the 40 videos annotated in the first crowdsourcing study). As before, we ask the crowdworkers to choose the class that best fits their assumption of the uploader intent for the video, from among the nine uploader intent classes (i.e., the first column of Table I). In contrast to the previous crowdsourcing HIT, we do not give the workers the possibility to suggest additional classes. We applied standard quality control practices, combining automatic analysis of completeness and consistency, with manual checks. Each video was annotated by three workers. In total, 621 unique workers participated in the video annotation process. The inter-annotator agreement, calculated with Fleiss kappa, surpassed the generally accepted level of 0.7 [28] with an average kappa of 0.791. The final annotation for each video was determined by majority

vote. In order to keep our work focused on cases that are intuitive (i.e., one of the criterion for the typology), we put the 523 videos for which no agreement was reached into a separate set for further investigation in future work. From this point on in this paper, we focus our investigations on the 3,052 videos for which inter-annotator agreement could be achieved.

Our analysis of patterns with which the study participants assigned uploader intent classes to videos starts with a study of the confusability between classes. Checking confusability allows us to verify the intuitiveness of the uploader intent typology, i.e., whether the classes are clear, recognizable and distinctive to humans attempting to interpret them. We found that workers had difficulty distinguishing between the classes that *convey knowledge* and that *explain something practical*. For this reason, we group these classes together with *illustrate* to form the class *Explaining* in the final typology, as can be seen in Table I. We acknowledge that a given video may provide a certain level of fit with a wide range of uploader intents. However, when discussing uploader intent and its prediction, similar to related investigations [1], [2], [8], we focus on the intent class that appears to provide the best fit with a video and considered to be its ‘dominant intent’.

Next, we look at the distribution of the videos over classes. Checking distribution allows us to verify the compactness of our typology, i.e., whether the classes are well balanced, and comparable in importance. In order to achieve balance we maintain *Entertaining* as its own class and group *Express affect or interests* with other classes under *Communicating* in the final typology, as shown in Table I. Note that *Entertaining* remains the largest class, consistent with observations in [29] on the dominance of entertainment content on YouTube.

Fig. 3 shows the distribution of videos in the five classes of the final uploader intent typology (black bar). In order to arrive at annotations of our 3,052 video dataset that use the five-class final uploader intent typology, we collapse the nine-class annotations collected from the workers in the second, large-scale crowdsourcing study (cf. Table I). Fig. 3 gives statistics on annotator agreement for each of the five classes (light and dark gray bars). The annotated video set is used for the experiments carried out in the remainder of the paper. The discovery of our uploader intent typology lets us positively answer *RQ1*.

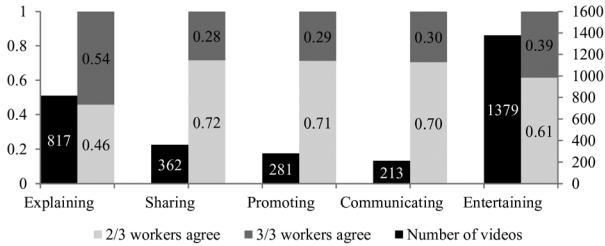


Fig. 3. Crowdsourcing-based uploader intent annotations and distribution of videos belonging to each uploader intent class (black) along with their inter-class agreement of workers (light and dark gray).

IV. MULTIMODAL VIDEO UPLOADER INTENT FEATURES

In this section, we address *RQ2*, ‘Are features extracted from textual and visual data associated with videos indicative of uploader intent?’ We analyze the importance of these features for automatic uploader intent prediction. This investigation is carried out on a development set of 250 videos, 50 from each intent class, randomly selected from our intent-annotated video collection.

A. Feature Extraction

Each video is represented by a set of multimodal features. In this, to the best of our knowledge, first investigation of uploader intent prediction, we make use of features extracted both from the *textual metadata* associated with a video as well as from its *visual channel* that have been proven useful in similar tasks. Textual features are derived from a video’s title ($T[Ti]$), description ($T[De]$), tags ($T[Ta]$) and the concatenation of all three ($T[TiTaDe]$). We arrive at the vocabulary by carrying out stop-word removal and stemming, and then calculating tf-idf.

We believe that useful, intent-sensitive visual variation of videos in different uploader intent classes can be encoded only with the appropriate feature choice. To this end, we extract state-of-the-art visual features from videos that have been frequently used to perform high-level visual recognition tasks similar to ours [26], [30] in unconstrained consumer-uploaded videos. We aim to investigate the usefulness of other visual features (e.g., as applied in [31]) as well as additional modalities (e.g., audio features as briefly investigated in [1]) in future work. We extract our features from the three keyframes provided by YouTube which we have downloaded for each video in our dataset. To capture motion-related information, we clip three seconds of contiguous frames around the keyframes to capture their immediate temporal past and future. We refer to these segments as *shots*. Note that this use of the word ‘shot’ does not necessarily match the standard definition of a shot in video content analysis [32].

We extract both low-level appearance and motion features, since they are often observed to capture complementary information [30]. For appearance-based features, a corner detector [33] is applied on keyframes and the detected corners (informative patches) are described using the Scale Invariant Feature Transform [33] (SIFT) algorithm. Here we use a 3-channel variant [34] of the traditional gray-SIFT descriptor [33] to also capture relevant color information. To extract motion information from shots, we employ the approach proposed in [35], in which corners are detected in the beginning frame of the shot and then tracked across the subsequent frames. These corners

are typically associated with objects in a video. By analyzing the velocity of the corner regions, we can select trajectories with strong motion as characteristic enough to represent a video. These trajectories are then described using Histogram of Orientated Gradient (HOG) and Motion Boundary Histogram (MBH). HOG captures appearance changes along a particular trajectory, while MBH captures motion related facets, such as velocity or acceleration. We perform hierarchical K-means clustering to generate vocabularies of visual words of different sizes ($K = 500, 1K, 2K, 5K, 10K$) for both appearance and motion features. Once vocabularies are obtained, features from each keyframe or a shot are quantized into histograms whose bins correspond to the visual words in the vocabularies. In other words, we generate global bag-of-X (X corresponding to either SIFT or HOG+MBH) representations for keyframes or shots ($V[SIFT]$ and $V[HM]$).

In addition to these features, we use mid-level semantic features derived from the visual channel that are designed to capture topical and affective aspects of video. We do not expect a strong correlation between topic, affect and uploader intent. However, we use these features in order to exploit any partial correlations that may exist. We extract topic-related mid-level representations from keyframes using ObjectBank [10], which contains 177 pre-trained generic detectors for concepts such as *sailboat* or *water*. We extract affect-related mid-level representations from keyframes using SentiBank [14], a large-scale classifier library containing concepts correlated with user expressions of sentiment associated with images. We apply the mid-level classifiers to each keyframe of each video in our dataset and represent each video with one feature vector from each concept set ($V[OB]$ and $V[SB]$). The components of the vectors are the confidence scores of the individual concepts averaged over the three keyframes.

B. Feature Analysis

We now turn to an investigation of the contribution of different features for uploader intent prediction.

Textual features: Fig. 4 presents term clouds of the 15 most important terms associated with each uploader intent class in the development data, derived from the $T[TiTaDe]$ feature. The term cloud suggests that textual terms are effective features, since top-weighted terms for each class are plausibly descriptive of the uploader intent of that class. For example, *Explaining* videos are correlated with terms such as *howto*, *tutorial*, *lesson*, *Promoting* with *trailer*, *commercial*, *product*, and *Entertaining* with *performance*, *song* and *episode*. *Sharing* and *Communicating* correlate with intent-related terms such as *event* or *travel*, but also with topic-dependent terms such as *obama*, *japan*, *lebron* or *dunk*.

Visual features: A priori, shots and keyframes can be expected to have wide visual variation with a single uploader intent class. In addition, many of the indicative terms in a potential vocabulary (e.g., *howto*, *product*, or *song*) would not have consistent visual patterns within videos since these concepts are highly abstract. While previous work [1], [2] dealing with user intent in video search confirms this expectation, it has also found that the visual channel can still serve as a weak indicator for prediction. In this paper, we also aim to exploit the visual channel as a weak predictor. For videos not having any textual metadata



Fig. 4. Term clouds of the 15 most important terms associated with each uploader intent class in the development data, derived from the T[TiTaDe] feature. The terms are extracted after stemming. The visual weight reflects that tf-idf value of each term: (a) *Explaining*; (b) *Sharing*; (c) *Promoting*; (d) *Communicating*; and (e) *Entertaining*.

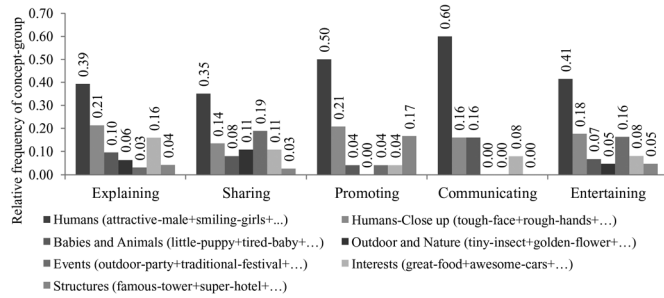


Fig. 5. Relative frequency of each group of semantically related concepts per uploader intent class along with two example concepts per group.

associated with them, classifiers operating on features solely extracted from the visual channel will be particularly helpful.

Mid-level visual representation: Next, we investigate whether particular concepts are indicative of uploader intent, i.e., whether videos associated with one intent class share commonalities in their semantic visual features. We focus our analysis on SentiBank concepts ($V[SB]$), since they offer insight of the visual channel including, but also going beyond topical content. For this reason, we anticipate that these features are particularly useful for uploader intent classification.

Naïvely, we could investigate the importance of each concept separately for each uploader intent class by, for example, calculating the tf-idf distribution of concepts per class. However, useful regularities are not necessarily manifested at the level of individual concepts, which may occur relatively infrequently. For this reason, we investigate *semantic groups of concepts* and how often they occur in videos associated with our uploader intent classes. In this analysis, we represent each video in our development set using only concepts having a certain minimum confidence of being present in the video. Filtering concepts by confidence score has been successfully applied in previous work to remove noise [11]. The threshold used to assess the confidence is set experimentally to 0.55. We use the filtered concept representation of each video, and build clusters by applying LDA to automatically discover groups of concepts appearing collectively in videos. For each group discovered in this way, we inspect its top-ranked, commonly appearing concepts and assign it a name reflecting its semantics. Then, for each concept group, we count the number of videos in each uploader intent class containing all concepts of that concept group.

Fig. 5 presents the relative frequency of the groups of semantically-related concepts over each uploader intent class. We observe that the groups of semantically-related concepts do indeed contribute differently to particular uploader intent classes. For example, *Explaining* and *Sharing* typically contain concept

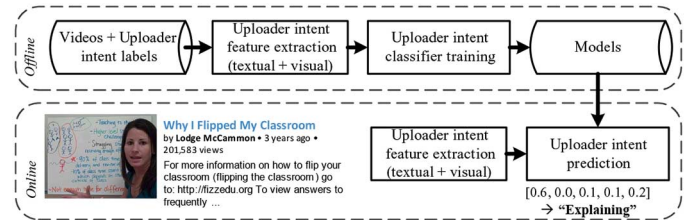


Fig. 6. Visualization of our uploader intent classification approach: We extract our defined features from uploader intent-labeled videos and train classifiers in an offline processing step. Our classifiers then perform automatic prediction of uploader intent classes for unseen videos in an online step.

groups related to *humans* (0.39, 0.35), but differ in *close-ups of humans* (0.21 vs. 0.14), *Interests* (0.16 vs. 0.11), *Events* (0.03 vs. 0.19) and *Outdoor and Nature* (0.06 vs. 0.11), groups of concepts typically related to these particular uploader intents. The observations in this section lead us to expect that multimodal, semantic features extracted from video can yield patterns related to uploader intent prediction, providing a positive answer to *RQ2*.

V. VIDEO UPLOADER INTENT PREDICTION

We use the features developed in the previous section to address *RQ3* ‘Can the class of uploader intent of a video be predicted automatically?’ Our classification approach, which follows a conventional architecture, is visualized in Fig. 6.

A. Experimental Setup

Our experiments are carried out on a test set containing 2,802 uploader intent-annotated videos (i.e., our 3,052 dataset minus the 250 video development set), with 767 video belonging to UI_{EX} , 312 to UI_{SH} , 231 to UI_{PR} , 163 to UI_{CO} and 1,329 to UI_{EN} (i.e., the five uploader intent categories in Table I). The URLs of these videos along with their uploader intent annotations are publicly available.²

Training, Prediction and Evaluation: We use a linear SVM [36] and train individual classifiers for each intent class and for each type of feature, and then combine their decisions. While training classifiers for one particular class, videos from remaining classes are used as negative samples. The optimal classifier parameters are obtained using a coarse-grid search on our development set. To fuse responses of classifiers trained from different representations, we use a simple multiplicative confidence fusion scheme [37]. We use 5-fold cross validation. We report our results in terms of accuracy, F-measure (FM)—the harmonic mean between precision and recall—and Weighted F-measure (WFM)—FM weighted by the test data

²[Online]. Available: <http://goo.gl/OKbHto>

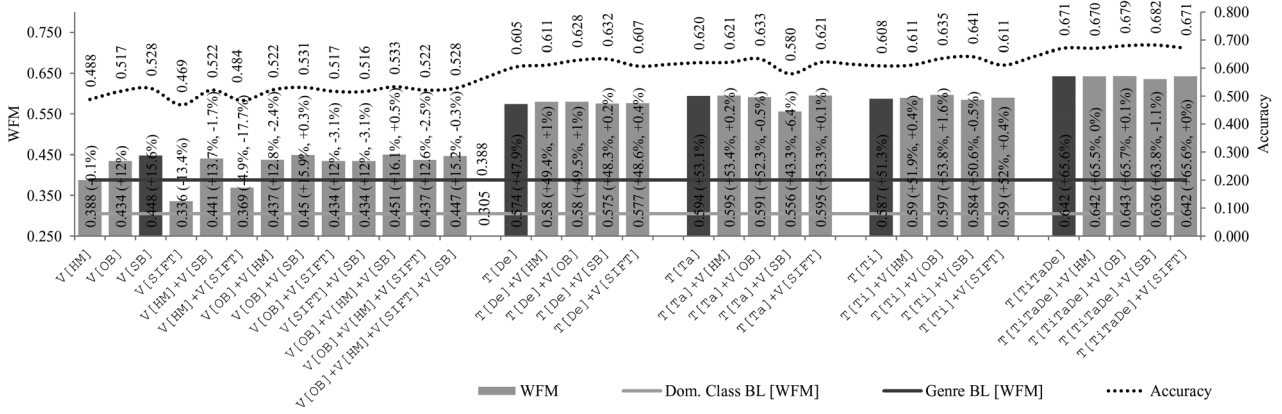


Fig. 7. Performance overview of our uploader intent classifiers presented in five groups: visual-only and four types of text (descriptions, tags, titles and their concatenation) plus visual combinations. Each bar in each group presents the performance achieved by methods trained on a particular feature set in terms of WFM. Dark gray bars present the best performance of approaches trained on a single, non-fused feature set. In parentheses, we report two comparisons: The first number reports the relative improvement of a bar’s presented performance over the best baseline (i.e., the genre baseline) and the second number reports the improvement over the best performance of approaches trained on a single, non-fused feature set (i.e., the performance presented by dark gray bars).

class size (i.e., the number of videos belonging to each class). Improvement between results is expressed on a relative scale.

Baselines: We compare our approach against two carefully chosen baselines. First, we use a standard dominant class baseline, which reflects a scenario in which all instances are automatically classified in the largest class (i.e., UI_{EN}), and serves as a sanity check. Second, we implement a baseline that assigns a vector of *YouTube genre categories* to videos, rather than our multimodal features. Genre is not a monolithic notion, but rather combines aspects of style, form and topic [9]. Since these are aspects also associated with uploader intent, it is reasonable to expect that genre categories are highly indicative of intent. For example, one of the *YouTube categories* is ‘Entertainment’, which, upon first consideration, seems like it could be synonymous with the uploader intent class *Entertaining*. Further, genre categories represent a competitive baseline, since genre information is available on YouTube and can be exploited without computationally expensive visual processing. We build a vector of all genre categories. The components corresponding to the genres assigned to a given video are set to one in the vector representing that video, and the remaining components are zero.

B. Experimental Results and Analysis

Our results are organized into five groups of classification approaches, depicted from left to right in Fig. 7. We observe that the genre baseline (dark gray line; 0.388) significantly outperforms the dominant class baseline (light gray line; 0.305). The first, leftmost group contains approaches exploiting only low- and mid-level visual features and their combinations. For non-fused feature sets, we observe that both mid-level content representations outperform the baselines: $V[SB]$ features (0.448) provide the best performance (+15.6% over the genre category baseline), and $V[OB]$ (0.434) features achieve the second-best performance (+12%). The best overall performance attained by visual-only features is provided by the late-fusion approach combining $V[OB] + V[HM] + V[SB]$ features (0.451) with a significant improvement of 16.1% over the genre category baseline, and a non-significant improvement of 0.5% over the $V[SB]$ approach.

It can be seen that these performance improvements are limited compared to approaches using text-based features (cf. the other four groups in Fig. 7). However, our visual-only-feature-based classifiers exploit the intent-aware visual signal in the data that is expected to be quite weak, as discussed in Section IV. The relatively good performance of the visual-only-based classifier using $V[SB]$ features, which are designed to capture visual semantics beyond topic, provides evidence that uploader intent also transcends topic. Although visual features provide only limited evidence for uploader intent, certain visual features carry more valuable information for uploader intent prediction than genre or conventional topically-related visual features do. In other words, almost all combinations based on visual features outperform the strong genre baseline, which is a useful and encouraging insight that can be worthwhile to steer the direction of research in automatic visual feature-based uploader intent analysis in future work. We also observe that mid-level representations outperform the low-level-feature-based approaches $V[HM]$ (0.388) and $V[SIFT]$ (0.336). The ability of certain visual features to outperform a competitive baseline is of critical importance for videos for which the uploader provides little or no textual metadata.

The remaining four feature groups (i.e., the four rightmost groups) present approaches involving our textual features and their combination with visual features. All approaches outperform the best baseline and visual-only-feature-based approaches. We observe that $T[TiTaDe]$ features (0.642) provide the best text-only approach (outperforming the genre baseline by 65.6%) and perform better than $T[Ta]$ (0.594), $T[Ti]$ (0.587) and $T[De]$ (0.574) approaches. These results suggest that our classifiers pick up particularly well on patterns in textual data that are truly uploader intent-specific and implicitly capture uploader intent. Based on the performance achieved by the individual text-only features, we observe that tags carry the most uploader intent-related information and that a video’s title is more indicative of uploader intent than its description. These observations are interesting, since they reveal that users apparently choose tags and titles related to the reason why they upload videos.





		Best Visual UI _{EX}	Best Textual UI _{PR}	Best Combined UI _{PR}
(1)	 <p>Personality Quiz Tags: beaverbunch, beaver, bunch, proud2badork, tuesday, GLBT, LGBT, lesbian, gay, bisexual, trans, community, personality, feelings</p>	✗	✗	✗
(2)	 <p>Wrestling 101: Takedowns, Referee's Position, Escape, Reversal, Scoring, Locked Hands Tags: okstate, osu, oklahoma, state, cowboys, ncaa, college, wrestling, amateur, high, school, olympics, usa, gold, medalist</p>	✓	✗	✗
(3)	 <p>Trailer : On the Road Official Trailer #1 (2012) - Amy Adams, Sam Riley Movie HD Tags: On the Road trailer, On the Road movie, On the Road 2012, On the Road HD, HD, 2012, Kristen Stewart, Amy Adams, Kirsten Dunst</p>	✗	✓	✗
(4)	 <p>Forza Motorsport 2: 1998 Toyota VellSide Supra Fortune 99 (99/346) Tags: Forza, Motorsport, Game</p>	✗	✗	✓

Fig. 8. Example videos and their uploader intent predictions [Source: YouTube. Video IDs: (1) BeWynYzUVrg, (2) eP1jRARJoVg, (3) obSHCIInYXFc, (4) Za1w493_4Zs].

Combining textual with visual features only slightly outperforms some text-only approaches. However, this improvement is not significant. The largest improvement of combined methods is achieved by combining T[De] with V[OB] (0.580) resulting in a 1% improvement over the text-only T[De] approach. The best overall performance is achieved by combining T[TiTaDe] with V[OB] (0.643). The relatively small improvement of combined methods is understandable: Our analysis in Section IV showed that text-based features carry more intent-specific information than visual features. Further, it also anticipates that the weak signal provided by visual features is helpful for a very limited number of cases where text-only-based approaches deliver wrong predictions (i.e., on average, 75 more videos get correctly classified using combined approaches).

In order to gain a better understanding of the strengths and weaknesses of our approach, we discuss the predictions made by our classifiers for four example videos (cf. Fig. 8). These videos were chosen because our predictions were most different from the prediction results obtained by the genre baseline. Video 1 (uploader intent is *Communicating*, UI_{CO}) is about a personality quiz targeted at the YouTube community, and it was confused by all classifiers. The visual classifier most probably made its decision because of the dominance of a person in the keyframes, which was not corrected by other approaches due to the lack of decisive textual metadata. Video 2 explains the basics of wrestling (uploader intent is *Explaining*, UI_{EX}) and was correctly classified by our visual classifier presumably due to the presence of particular semantic concept groups. However, it was confused by the other approaches. Tags such as *cowboys* or *olympics* apparently hinted to the classifier that the video has an UI_{EN} background. Video 3 is a trailer of a movie (uploader intent is *Promoting*, UI_{PR}) and was correctly classified by our text-only classifier presumably based on patterns in metadata such as the combination of *trailer* and *movie*. Video 4 is a recording of a video game (uploader intent is *Entertaining*, UI_{EN}) and was only correctly classified by the classifier exploiting both visual features and text, i.e., complementary features of specific semantic visual concept groups and keywords such as *game* and *motorsport* resulted in correct predictions. Our experimental results and analysis demonstrate that automatic classifiers can infer video uploader intent, and allow us to answer RQ3 positively.

VI. APPLICATIONS OF VIDEO UPLOADER INTENT

In this section, we turn to the question of the usefulness of video uploader intent in applications. We investigate the contribution of uploader intent to two prediction tasks already introduced in Section I, namely, *reach* (in Section VI-A) and *impact* (i.e., likecount, viewcount and rating) of videos (in Section VI-B). In a final small-scale user study, we explore the relationship between the intent of a user searching for video, and a user uploading a video (in Section VI-C), which is critical if inferred uploader intent is to be used to improve video search results, e.g., via reranking.

Features: To investigate *reach* and *impact*, we perform experiments making use of both oracle (ground truth uploader intent classes assigned in the annotation process) as well as predicted (scores predicted by our classifiers) uploader intent. Using oracle information allows us to investigate whether uploader intent, theoretically, makes a contribution to the task. Using inferred information allows us to investigate the ability of our uploader intent classification approach (Section V) to contribute to real-world applications.

In both prediction tasks, a video is represented by a vector whose components indicate the confidence of how well the video satisfies a particular intent. For oracle uploader intent, we refer to this vector as \mathbf{U}_O and calculate the fraction of how many workers voted for a particular uploader intent class in our crowdsourcing annotation process. For predicted uploader intent, we use the classifier outputs from our overall best-performing uploader intent classifier (i.e., T[TiTaDe] + V[OB]), subsequently referred to as \mathbf{U}_{T+V} . To evaluate performance for videos with no textual metadata, we experiment with outputs generated by our best-performing visual-only intent classifier (i.e., V[OB] + V[HM] + V[SB])— \mathbf{U}_V . These vectors have the following structure: [UI_{EX}, UI_{SH}, UI_{PR}, UI_{CO}, UI_{EN}].

Baselines and Evaluation: For each task, we compare our approach with the dominant class baseline and also with the genre-category baseline. We also evaluate our approaches against the *Text baseline* where each video is represented by a tf-idf vector generated from its title, tags and description. Previous work [1], [22], but also our feature analysis in Section IV-B, suggests that textual metadata carries valuable information for various prediction approaches. We are interested to determine whether uploader intent achieves comparable or superior performance for predictions. Due to the insights obtained in the experiments presented above, we do not experiment with any baseline approaches exploiting visual features, as they are not expected to outperform the textual baseline. We train and evaluate our classifiers for the prediction tasks using the same fusion approach as our uploader intent classifier. Statistical significant tests were performed using the Wilcoxon signed-rank test ($\diamond : p < 0.01, \Delta : p < 0.05$).

A. Reach: Video Uploader Intent and Target Audience

Here, we address research question RQ4 on *reach*, ‘Can the uploader intent class be used to predict the size of the audience at which the uploader targeted the video?’ Motivation for investigation of *reach* prediction lies in the assumption that being able to automatically predict the size of the target audience of a video, at the moment that the video is uploaded, could be

TABLE II
TARGET AUDIENCE CLASSES DISCOVERED IN OUR CODING PROCESS

Target audience	Description
<i>Personal</i> (TA_{PE})	person or close group of people the uploader knows personally (e.g., family or friends).
<i>Social</i> (TA_{SO})	group of people sharing the uploader’s interests, but who the uploader does not necessarily know personally.
<i>Public</i> (TA_{PU})	wide audience and not for a particular group of people.

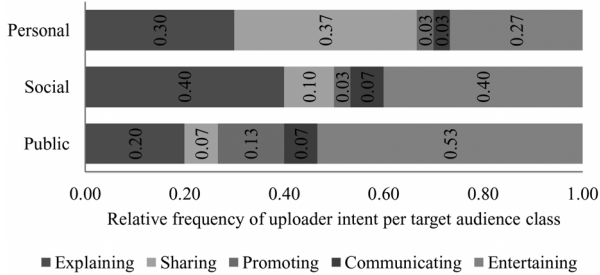


Fig. 9. Correlation between uploader intent and target audience classes.

helpful in a number of applications. These include improved matching of videos with advertisement, providing uploaders with recommendations for metadata, i.e., which particular keywords should be contained in titles, tags and descriptions, and improving search results ranking and filtering.

In order to define size-related target audience classes for investigation, we carry out a manual coding process similar to the one applied to discover uploader intent classes (cf. Section III-A). We manually inspect the title, description and tags of the 250 videos in our development set and iteratively define a set of three target audiences: *personal*, *social*, and *public*, presented in Table II. We used our crowdsourcing annotation process (cf. Section III-D) to generate target audience ground truth concurrently with uploader intent class ground truth. For the 2,802 uploader intent-annotated videos in our dataset, 2,684 videos were annotated with target audiences (for 118 videos no agreement could be found) with a distribution of 85 videos belonging to TA_{PE} , 841 to TA_{SO} and 1,758 to TA_{PU} (kappa: 0.814). The reason for the relatively low number of videos belonging to the TA_{PE} class could be that some videos that would belong to this class are not publicly shared. However, note that we do not claim that we study all *existing* personal videos. Instead, we believe that uploaders who publicly share personal videos do not mind if these videos are available to a larger audience. Therefore, we believe that our dataset is a representative sample of videos in terms of our target audience classes as they exist for publicly-shared videos and satisfactory for the proof-of-concept discussed here.

In order to investigate the relationship between uploader intent class and target audience class, we randomly sample 90 videos from our dataset, i.e., 30 from each target-audience class. For each target-audience class, we calculate the relative frequency of videos belonging to different uploader intent classes (cf. Fig. 9). We observe that the majority of videos uploaded for target audience *Personal* have the goal to share (37%) or to explain (30%); *Social* target audience videos typically explain (40%) or entertain (40%). The majority of videos uploaded for a *Public* audience typically have the intent to entertain (53%), but

TABLE III
OVERVIEW OF OUR TARGET AUDIENCE PREDICTION RESULTS

Method	FM TA_{PE}	FM TA_{SO}	FM TA_{PU}	WFM	Acc.
Dominant class baseline	0.000	0.000	0.652	0.315	0.484
YouTube genre categories baseline	0.000	0.606	0.488	0.529	0.546
Text baseline	0.088	0.575	0.554	0.549	0.555
Oracle uploader intent U_O	0.308	0.603	0.620	0.601 ◇	0.606
Pred. uploader intent U_{T+V}	0.222	0.565	0.610	0.575 △	0.582
Pred. uploader intent U_V	0.000	0.159	0.619	0.427	0.539

it is also understandable that videos uploaded to promote (13%) are typically related with this target audience to maximize the number of views.

Next, we analyze the potential of automatic methods predicting the target audience for videos using uploader intent. For our experiments, we remove the 90 videos previously selected for our manual investigation and restrict the number of videos in the largest class (i.e., *Public*) by randomly sampling 811 videos from it to have a more balanced distribution among our three target audience classes. This results in an experimental dataset containing 1,677 videos and a distribution of 55 videos belonging to TA_{PE} , 811 to TA_{SO} and 811 to TA_{PU} . An overview of our experimental results is presented in Table III. The *text baseline* is the strongest baseline approach (0.549) and outperforms the genre-category baseline (0.529) and the dominant class baseline (0.315). These findings are interesting since they show that information extracted from the textual metadata associated with videos provides a comparably strong signal for target audience prediction and that these features provide more evidence for target audience than the genre categories of videos. The solid performance of these features also provides evidence that target audience is highly correlated with uploader intent, since these feature sets already delivered good prediction results for uploader intent prediction. In addition, it is the only baseline approach which correctly predicts a limited amount of *Personal* videos (0.088).

Features U_V (0.427) solely outperform the dominant class baseline and do not achieve performance comparable to our best baseline approach. However, this result suggests that U_V still provides a weak indicator for target audience prediction, which can be exploited for videos not having genre categories or textual metadata assigned. Our fully automatic target audience prediction approach exploiting U_{T+V} features (0.575) statistically significantly outperforms the best baseline approach (+4.7%). Further, the approach achieves performance comparable with that of the overall best target audience prediction approach, which exploits oracle features U_O (0.601; +9.4%). This result confirms our observations that uploader intent provides more contextual information for prediction of the intended target audience, than textual metadata associated with videos. This point becomes particularly clear when observing the performance achieved for the *Personal* class (cf. 0.222 by U_{T+V} and 0.308 by U_O with 0.088 by the text baseline). Most importantly, approaches exploiting uploader intent significantly outperform the baseline approach exploiting genre (cf. 0.575 by U_{T+V} and 0.601 by U_O with 0.529 by the genre baseline). These findings let us positively answer *RQ4*.

TABLE IV
OVERVIEW OF OUR POPULARITY PREDICTION RESULTS

Method	MSE	MSE	MSE
	Viewcount	Likecount	Rating
YouTube genre category baseline	0.0238	0.0234	0.0877
Text baseline	0.0236	0.0230	0.0857
Oracle uploader intent \mathbf{U}_O	0.0180◇	0.0174◇	0.0773△
Pred. uploader intent \mathbf{U}_{T+V}	0.0212△	0.0211△	0.0829
Pred. uploader intent \mathbf{U}_V	0.0273	0.0285	0.1023

B. Impact: Video Uploader Intent and Popularity

Here, we address *RQ5* and investigate whether characteristics associated with popularity of videos can be predicted automatically using uploader intent. Similar to how target audience information can be exploited by search engines to derive more context from videos, popularity prediction can be used to determine whether uploaded videos will become viral, which can ultimately also be used by users who upload the video or exploited by search engines for faster video delivery, advertisement and related applications.

In order to create models for popularity prediction of videos given their uploader intent distribution, we rely on the popularity-related information we collected for our YouTube videos. For training and prediction, we use popularity values that are normalized by the period of time that has passed from the time videos were uploaded until our dataset was collected. We train individual Support Vector Regressors [36] for viewcount-, likecount- and rating-prediction. We select optimal regression parameters for each experiment using coarse-grid search on our development set and perform 5-fold cross validation during testing (cf. Section V-A). We report performance in terms of Mean Squared Error (MSE). Table IV contains an overview of our prediction performance.

We observe that the text baseline carries more value for popularity prediction than the genre baseline. Although the difference is not statistically significant, we note that this result suggests that textual metadata associated with a video may be more indicative of its popularity than its genre category. The predictors exploiting visual-only-based uploader intent information from \mathbf{U}_V features do not compete with the performance of the best baseline. However, we observe that our approach exploiting \mathbf{U}_O features significantly outperforms the best baseline for all popularity features. For viewcount (0.0180 vs. 0.0236) we achieve an improvement of 23.7%, for likecount (0.0174 vs. 0.0230) 24.4% and for rating (0.0773 vs. 0.0857) 9.8%. The performance of the predictors exploiting \mathbf{U}_{T+V} features is comparably good, i.e., in a live setting, we still achieve 10.2%, 8.3% and 3.3% performance improvement over the text baseline. Note that we do not claim that our approach performs better than previously proposed techniques (e.g., [20]). Instead, the purpose of this experiment is to investigate the connection between uploader intent and the popularity of videos. We answer *RQ5* positively, since our results suggest that uploader intent contributes to the prediction of popularity for videos.

C. Search: Video Uploader Intent and Query Search Intent

Here, we report on a final experiment that explores the connection between uploader intent and search intent, i.e., *RQ6*. As mentioned in the Introduction, video search is an important application for uploader intent. A positive correlation be-

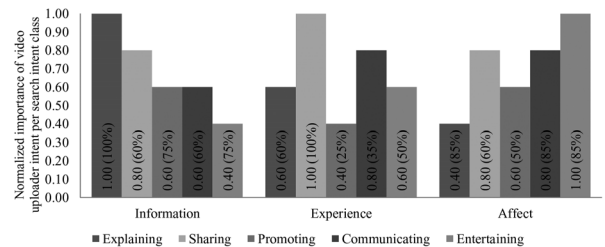


Fig. 10. Correlation between our discovered uploader intent classes and search intent classes from [1].

tween search and uploader intent could ultimately be exploited for novel, intent-aware search engine optimizations. The purpose of this validation experiment is to demonstrate the extent of the value of uploader intent in search. We look for clues that confirm that uploader intent (motivation for uploading video) is related to search intent (the reasons for which users are trying to find video) and leave actual intent-aware search results list optimizations (e.g., such as investigated in [2]) for future work. Specifically, we investigate which of the discovered uploader intent classes users expect in lists of search results that are relevant to a specific search intent.

For our investigations, we adopt an established video search intent typology of three basic intent classes [1]: *Information*—users aim to obtain declarative knowledge, i.e., obtain information, or performative knowledge, i.e., acquire a skill; *Experience*—users aim to have particular experiences of an actual person, place, entity, event etc.; and *Affect*—users aim to change their mood or affective state, i.e., be entertained. We carry out a single-task user study on AMT. The study is formulated abstractly, and is independent of actual queries or videos. Instead, it asks users about the connections that they expect between reasons why people search for videos, and reasons why they upload them.

We present workers with the three query search intent classes and our five uploader intent classes accompanied by descriptions and examples. For each search intent class, we ask workers to rank uploader intent classes according to importance. In our setup, each rank must be unique for each search intent class, i.e., all ranks 1 through 5 must be selected and no duplicate ranks are allowed.

In total, 20 workers participated in our user study—a representative sample size for investigations following qualitative principles [38]. For each search intent-uploader intent-pair, we apply inter-annotator agreement using majority voting to fuse ranks provided by our workers and normalize these ranks to obtain the final importance scores of each uploader intent class. Fig. 10 presents these importance scores accompanied by the agreement percentage among workers (in parentheses) for each uploader intent class. Note that due to the fusion process of scores, two or more uploader intent classes may receive the same importance score for a search intent class.

We observe evidence that videos that would fulfill a certain search intent, are related to certain classes of uploader intent: *Information* correlates mostly with uploader intent *Explaining*; *Experience* with *Sharing*; and *Affect* with *Entertaining*. These results provide a confirmation of the potential of uploader intent for improving video search.

We observe that the agreement among workers varies by search intent: For *Information*, *Experience*, and *Affect*, respectively, 74%, 54%, and 73% of workers agree on the rank they provided. The comparably low agreement for search intent *Experience* results from the fact that workers disagree on which uploader intents are only *partially* important for this search intent (e.g., low agreement for classes *Promoting* and *Communicating*). We discover that the agreement among workers for most-relevant and most-irrelevant uploader intents per search intent is typically high. For example, all workers agreed that for search intent *Information*, uploader intent class *Explaining* is most relevant and 15/20 workers (75%) agreed that class *Entertaining* is most irrelevant. This result suggests that the importance of particular uploader intents highly depends on the search intent of a query. This can be exploited by intent-aware optimization techniques, which should focus on properly reranking initial search results lists. Such optimization can be accomplished, for example, by applying reranking weights that are derived from the presented uploader intent-search intent correlation, eventually producing results lists that provide a better match between the query's search intent and the uploader intents covered in the top-ranked results. Our findings let us positively answer *RQ6*.

VII. CONCLUSION AND OUTLOOK

We have presented a novel approach that automatically infers the uploader intent of videos, i.e., the reason or purpose expressing why users upload videos to the Internet, and have shown its usefulness in three application scenarios. Our approach is based on a typology of uploader intent classes, and uses multimodal representations of videos to learn a classifier capable of automatically inferring intent.

The paper lays the groundwork for the future study of uploader intent for video. There are a number of fertile aspects of uploader intent that have not yet been thoroughly explored. First, further work is necessary to understand fully the contribution that can be made by content features, including audio features and a fuller range of visual features (e.g., such as those in [31]). Second, it is interesting to separate video production from the act of uploading. The two are not necessarily synonymous, and understanding their connection may provide additional useful intent-related information about video. Third, the development of new intent classes is interesting. We point out that video sharing technology itself appears to give rise to new goals for users. In the early days of Internet video, videos that 'went viral' achieved skyrocketing popularity by chance, rather than by the explicit intent of the uploader. Today, users are increasingly attempting to design videos with the sole intent of having them watched by millions, and possibly of earning income in the process. Finally, we aim to improve search results ranking by exploiting the discovered correlation between video uploader intents and query search intents. We believe that this correlation will help the improvement of intent-aware reranking models, which have only been briefly investigated [2].

Our work points to the conclusion that why users upload videos to the Internet should be studied on equal footing with the topic and affective impact of video. Video search engines will be able to reach their full potential only if we take as much

information as possible about videos into account, and this information naturally includes uploader intent.

REFERENCES

- [1] A. Hanjalic, C. Kofler, and M. Larson, "Intent and its discontents: The user at the wheel of the online video search engine," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 1239–1248.
- [2] C. Kofler, M. Larson, and A. Hanjalic, "Intent-aware video search result optimization," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1421–1433, Aug. 2014.
- [3] T. Kindberg, M. Spasojevic, R. Fleck, and A. Sellen, "The Ubiquitous Camera: An in-depth study of camera phone use," *IEEE Pervasive Comp.*, vol. 4, no. 2, pp. 42–50, Apr. 2005.
- [4] M. Lux and J. Huber, "Why did you record this video? An exploratory study on user intentions for video production," in *Proc. WIAMIS*, 2012, pp. 1–4.
- [5] M. Campanella and J. Hoonhout, "Understanding behaviors and needs for home videos," in *Proc. 22nd Brit. HCI Group Annu. Conf. People and Computers*, 2008, pp. 23–26.
- [6] N. Bornoe and L. Barkhuus, "Video microblogging: Your 12 seconds of fame," in *Proc. CHI '10 Extended Abstracts*, 2010, pp. 3325–3330.
- [7] N. Park, Y. Jung, and K. M. Lee, "Intention to upload video content on the internet: The role of social norms and ego-involvement," *Comput. Human Behavior*, vol. 27, no. 5, pp. 1996–2004, Sep. 2011.
- [8] M. Lux, C. Kofler, and O. Marques, "A classification scheme for user intentions in image search," in *Proc. 28th Int. Conf. Human Factors Computing Syst.*, Apr. 2010, p. 3913.
- [9] M. Larson *et al.*, "Automatic tagging and geotagging in video collections and communities," in *Proc. 1st ACM Int. Conf. Multimedia Retrieval*, 2011, pp. 1–8.
- [10] L.-J. Li, H. Su, E. P. Xing, and F.-F. Li, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. Neural Inf. Process. Syst.*, 2010, pp. 1378–1386.
- [11] S. Rudinac, M. Larson, and A. Hanjalic, "Leveraging visual concepts and query performance prediction for semantic-theme-based video retrieval," *Int. J. Multimedia Inf. Retrieval*, vol. 1, no. 4, pp. 263–280, 2012.
- [12] S. Schmiedeke, P. Kelm, and T. Sikora, "Cross-modal categorisation of user-generated video sequences," in *Proc. 2nd ACM Int. Conf. Multimedia Retrieval*, 2012, pp. 25:1–25:8.
- [13] J. Chen, Y. Cui, G. Ye, D. Liu, and S.-F. Chang, "Event-driven semantic concept discovery by exploiting weakly tagged internet images," in *Proc. Int. Conf. Multimedia Retrieval*, 2014, pp. 1:1–1:8.
- [14] D. Borth, T. Chen, R. Ji, and S.-F. Chang, "SentiBank: Large-scale ontology and classifiers for detecting sentiment and emotions in visual content," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 459–460.
- [15] A. Habibian and C. G. Snoek, "Video2Sentence and vice versa," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 419–420.
- [16] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. San Pedro, "How useful are your comments?: Analyzing and predicting YouTube comments and comment ratings," in *Proc. 19th Int. Conf. World Wide Web.*, 2010, pp. 891–900.
- [17] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I Tube, You Tube, Everybody Tubes: Analyzing the world's largest user generated content video system," in *Proc. 7th ACM SIGCOMM Conf. Internet Meas.*, 2007, pp. 1–14.
- [18] F. Figueiredo, F. Benevenuto, and J. M. Almeida, "The Tube over time: Characterizing popularity growth of Youtube videos," in *Proc. 4th ACM Int. Conf. Web Search Data Mining*, 2011, pp. 745–754.
- [19] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Commun. ACM*, vol. 53, no. 8, pp. 80–88, Aug. 2010.
- [20] H. Pinto, J. M. Almeida, and M. A. Gonçalves, "Using early view patterns to predict the popularity of youtube videos," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, 2013, pp. 365–374.
- [21] X. Cheng, C. Dale, and J. Liu, "Understanding the characteristics of internet short video sharing: YouTube as a case study," *CoRR*, vol. abs/0707.3670, Jul. 2007 [Online]. Available: <http://arxiv.org/abs/0707.3670>
- [22] I. Weber, V. R. K. Garimella, and E. Borra, "Inferring audience partisanship for YouTube videos," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 43–44.
- [23] A. Abisheva, V. R. K. Garimella, D. Garcia, and I. Weber, "Who watches (and shares) what on youtube? and when?: Using twitter to understand youtube viewership," in *Proc. 7th ACM Int. Conf. Web Search Data Mining*, 2014, pp. 593–602.

- [24] M. Naaman, H. Becker, and L. Gravano, "Hip and trendy: Characterizing emerging trends on Twitter," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 62, no. 5, pp. 902–918, 2011.
- [25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
- [26] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui, "Consumer video understanding: A benchmark database and an evaluation of human and machine performance," in *Proc. 1st ACM Int. Conf. Multimedia Retrieval*, 2011, pp. 29:1–29:8.
- [27] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White, "Evaluating implicit measures to improve web search," *ACM Trans. Inf. Syst.*, vol. 23, no. 2, pp. 147–168, Apr. 2005.
- [28] M. Warrens, "Inequalities between multi-rater kappas," *Adva. Data Anal. Classification*, vol. 4, no. 4, pp. 271–286, 2010.
- [29] X. Cheng, J. Liu, and C. Dale, "Understanding the characteristics of internet short video sharing: A YouTube-based measurement study," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1184–1194, Aug. 2013.
- [30] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah, "High-level event recognition in unconstrained videos," *Int. J. Multimedia Inform. Retrieval*, vol. 2, no. 2, pp. 73–101, 2013.
- [31] M. Riegler, M. Larson, M. Lux, and C. Kofler, "How 'how' reflects what's what: Content-based exploitation of how users frame social images," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 397–406.
- [32] A. Hanjalic, "Shot-boundary detection: Unraveled and resolved?," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 2, pp. 90–105, Feb. 2002.
- [33] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. Int. Conf. Comput. Vis.*, 1999, vol. 2, p. 1150.
- [34] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010.
- [35] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 3169–3176.
- [36] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008.
- [37] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.
- [38] M. Williams, "Interpretivism and generalisation," *Sociology*, vol. 34, no. 2, pp. 209–224, 2000.



Christoph Kofler received the B.Sc. and M.Sc. degrees in computer science from Klagenfurt University, Klagenfurt, Austria, and is currently working toward the Ph.D. degree at the Delft University of Technology, Delft, The Netherlands.

He has held positions at Microsoft Research, Beijing, China, Columbia University, NY, USA and Google, NY, USA. His research interests include multimedia information retrieval with a focus on video search intent inference and its impact on search optimization.

Mr. Kofler is the recipient of the Google Doctoral Fellowship.



Subhabrata Bhattacharya (M'14) received the Ph.D. degree in computer engineering from the University of Central Florida, Orlando, FL, USA, in 2013.

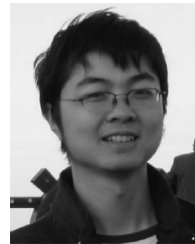
He is a Research Scientist with Imaging and Computer Vision, Siemens Corporation, Corporate Research, Princeton, NJ, USA. He was previously a Postdoctoral Researcher with the Digital Video and Multimedia Lab, Columbia University, New York, NY, USA. His research interests are in computer vision, including content-based video analysis and aesthetic understanding of images and videos.



Martha Larson (M'09) received the B.S. in mathematics from the University of Wisconsin, Madison, WI, USA, and the M.A. and Ph.D. degrees in theoretical linguistics from Cornell University, Ithaca, NY, USA.

She is currently an Assistant Professor in multimedia computing with the Delft University of Technology, Delft, The Netherlands. She is a Cofounder of the MediaEval Multimedia Benchmark. Her research interests include multimedia information retrieval and the use of crowdsourcing for studying and improving information systems.

Dr. Larson is a frequent member of organization and program committees of workshops and conferences, and is currently an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA.



Tao Chen received the B.S. degree in fundamental science and Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2005 and 2011, respectively.

He is currently a Postdoctoral Researcher with the Digital Video and Multimedia Lab, Columbia University, New York, NY, USA. His research interests include multimedia, computer graphics and computer vision.

Dr. Chen was the recipient of the Netexplorateur Internet Invention Award of the World in 2010 and the China Computer Federation Best Dissertation Award in 2011.



Alan Hanjalic (SM'08) received the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, and the Diplom-Ingenieur (Dipl.-Ing.) degree from the Friedrich-Alexander University, Erlangen, Germany, both in electrical engineering.

He is currently a Professor of computer science and head of the Multimedia Computing Group, Delft University of Technology, Delft, The Netherlands. His research focus is on multimedia information retrieval and recommender systems.

Prof. Hanjalic is Chair of the Steering Committee of the IEEE TRANSACTIONS ON MULTIMEDIA, an Associate Editor-in-Chief of the *IEEE MultiMedia Magazine*, and a member of the editorial boards of the *ACM Transactions in Multimedia*, the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, and the *International Journal of Multimedia Information Retrieval*.



Shih-Fu Chang (M'88–F'04) is the Richard Dicker Professor and Senior Vice Dean at Columbia University Engineering School, New York, NY, USA. His research is focused on multimedia information retrieval, computer vision, signal processing, and machine learning.

Prof. Chang is a Fellow of the American Association for the Advancement of Science. He served as the Editor-in-Chief of the *IEEE Signal Processing Magazine* from 2006 to 2008. He was the recipient of the IEEE Signal Processing Society Technical Achievement Award, the ACM Multimedia SIG Technical Achievement Award, the IEEE Kiyo Tomiyasu Award, and the IBM Faculty Award.