YouTube Timed Metadata Enrichment Using a Collaborative Approach

José Pedro Pinto[1] and Paula Viana[1,2](✉)

[1] INESC TEC, Porto, Portugal
{jppinto,pviana}@inesctec.pt
[2] School of Engineering, Polytechnic of Porto, Porto, Portugal

Abstract. Although the growth of video content in online platforms has been happening for some time, searching and browsing these assets is still very inefficient as rich contextual data that describes the content is still not available. Furthermore, any available descriptions are, usually, not linked to timed moments of content. In this paper, we present an approach for making social web videos available on YouTube more accessible, searchable and navigable. By using the concept of crowdsourcing to collect the metadata, our proposal can contribute to easily enhance content uploaded in the YouTube platform. Metadata, collected as a collaborative annotation game, is added to the content as time-based information in the form of descriptions and captions using the YouTube API. This contributes for enriching video content and enabling navigation through temporal links.

Keywords: Video tagging · Video retrieval · Crowdsourcing
Multimedia content annotation · Gamification · Social media · YouTube

1 Introduction

Online video platforms have provided the ground for this type of content to become widely used and a source of information and entertainment for millions of users. There are numerous video sharing websites such as Vimeo, YouTube and Dailymotion. Among those, YouTube is certainly the most popular, with videos shared amongst 1.3 billion of users across the whole world. With more than 300 h of video uploaded every minute, YouTube offers a novel kind of knowledge base for multimedia data. Apart from allowing users to upload and share their videos, it also encourages them to enrich the visual content with context information that includes tags, categories, title, etc. This process results in coupling massive amounts of content with user-generated metadata that greatly facilitates video retrieval and browsing by using text-based search engines.

However, the existing metadata is usually linked to the whole video and no timecoded annotations are available. Therefore, search performance and accuracy are reduced since users must watch the entire video to find parts of their interest – a tedious and time-consuming task. This lack of timecoded annotations will make some of the users to miss the chance to watch the intended scenes at a specific time [1].

K. Choroś et al. (Eds.): MISSI 2018, AISC 833, pp. 131–141, 2019.
https://doi.org/10.1007/978-3-319-98678-4_15

This drawback can be overcome by generating descriptions associated to specific points in the video. However, manually annotating video content is an expensive and time-consuming process. These almost incompatible aspects are the drivers for finding new methods that enable the creation of richly described video assets. Although some video annotation systems have been proposed, no solution has been yet provided for creating metadata to improve third party platforms like YouTube.

In this paper, we propose using a collaborative video annotation game platform to extend YouTube metadata with timecoded descriptions. Crowdsourced metadata created in the game produces tags of YouTube videos which are then exported back to YouTube as description and captions files in order to be indexed by YouTube's search engine. Our work will contribute to create better video content descriptions, which will enhance video searching results, as well as help users to quickly find scenes of their interest without having to watch the full stream.

2 Related Work

Commenting on videos is an approach for users to contribute with opinions and discuss some of the content on the video. Using YouTube comments to facilitate access and retrieval of online videos has already been exploited. On their proposal [2], authors describe a set of temporal transformations for multimedia content that allows end-users to create and share personalized timed-text comments that are combined chronologically. A survey confirms a better user experience when watching videos together with these timed related comments. However, this solution uses the deprecated Flash Player.

Based on social activities, especially user comments and weblog authoring, the work presented in [3] describes a mechanism that helps users to associate video scenes with user comments, to generate entries that quote video scenes, and to extract deep-content-related information about video contents for automatic annotation. This solution is made available as a standalone Windows application, not web-based. Moreover, it is too complex for a standard user and requires the user to fill a lot of specific metadata fields to annotate a simple video.

Annotations that include sentiment analysis and emotion modelling based on YouTube channel comments has been proposed by [4]. The solution developed uses gamification approaches to help on the collection of the information that is then used to enable content recommendation.

A video scene annotation method based on tag clouds has also been proposed [5]. Comments associated to a video and collected from YouTube are processed and a tag cloud is generated based on those comments. Based on the user clicks on a tag in the cloud while watching the video, the tag gets associated with the scene in the video. However, the presented web videos don't use HTML5 technologies and the set of available tags is restricted to the ones that are extracted from the comments.

A collaborative annotation system of social media that includes temporal duration of the scenes and uses ontological themes of the selected domain has been proposed in

[6] The application allows users to annotate content using free-text or following ontologically rules, with the objective of enhancing faster retrieval when browsing and searching for videos (specific scene, events, object, etc.).

Speech recognition has also been used to improve descriptions of online content. A framework for extracting relevant information from the audio track is exploited in [7]. Results show that the superimposition of relevant text and image-based information could be used for augmenting the viewing experience, as well as to give a full context-aware perception.

A browser extension that enables crowdsourcing of event detection in YouTube videos through a combination of textual, visual and behavior analysis techniques has also been proposed [8]. Based on the analysis of the visual content, it offers the user the choice to quickly jump to a specific shot in the video by clicking on a representative frame. The available metadata combines the one uploaded by the video owner such as title, description, tags, etc. and closed captions, which can be user-generated or auto-generated via speech-recognition. Interest-based event detection is achieved by counting clicks on shots. Although this seems promising, it requires installing a browser extension, what could be an obstacle for a significant number of people. Furthermore, as in the examples above, the produced metadata is not stored on YouTube and does not then contribute to enrich the platform's video content.

Besides the limitations already identified, it is worth mentioning that the most important drawback of these proposals is the fact that metadata is stored locally, and only local users who provided the annotations have access to this information. So, there is no solution for making videos enhanced with richer content information available for the full YouTube's community. Additionally, although user comments are quite popular in YouTube, they are also extremely controversial and usually acknowledged as very noise. Given that heaps of comments are continuously posted every day, the task of filtering good video related comments is not easy.

YouTube captions mechanism has proven to be a good method to increase views. An experiment has found an overall increase of 7.32% in views for captioned videos [9]. Professional services are even available for adding captions to YouTube videos [10]. However, the service is paid and only transcribes the audio.

In our proposal, we try to overcome the identified limitations by implementing mechanisms for the validation of descriptions provided by users and by making this validated metadata available to the full YouTube's community using the captions and description fields.

3 YouTube Video Platform

Social web videos have pervaded on the web, along with contextual information that describes the content, easing video browsing and searching. However, a great part of the online platforms provides only succinct and generic information, with no temporal pointers to specific happenings. There are numerous video sharing websites, such as Vimeo, YouTube, Dailymotion, etc., where people can share their ideas and thoughts by sharing videos.

The online video sharing website YouTube was originally created in February 2005 to help people share videos of personal or well-known events. It provides a forum for people to engage with video content across the world and acts as a distribution platform for content creators. YouTube content ranges from professional to amateur and the diversity of videos goes from TV clips to short videos with a variety of content types, such as tutorials, educational videos, music clips, video blogging, etc. The popularity of the platform is driven by the easiness of sharing and reproducing video content [11].

3.1 Search Engine

YouTube doesn't yet include the tools for automatic scene description. This means that the system depends on provided metadata and relevant information to help users to find content when searching for something. Therefore, uploaders should create useful and optimized metadata to have better chances for their video to be found. Information as title, tags, description, closed captions and user's reactions, are some of the useful information that YouTube search engine will use to present more precise results on a search query.

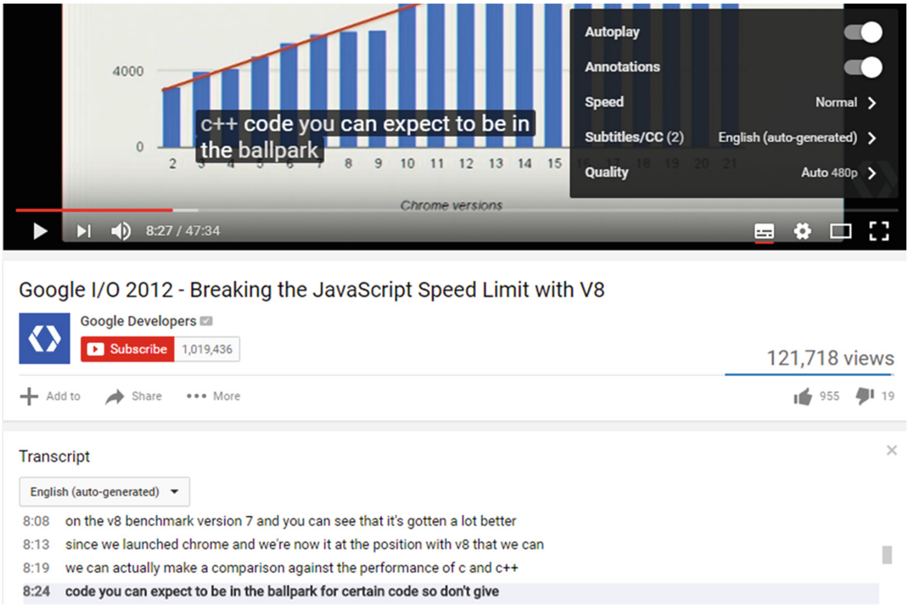
For the purpose of this paper, we will focus on a new approach of using the captions and description features to improve search for videos and, more important, search within videos on YouTube.

3.2 Captions Mechanisms

Under the video library management page, YouTube allows owners to upload their own closed captions file as subtitles to YouTube videos. A subtitle or closed caption file contains the text transcription of the audio stream and the time codes for when each line of text should be displayed. Some files also include position and style information enabling the customization of the presentation.

YouTube supports a vast list of captions' file formats. The most usual formats are SubRip (.srt), SubViewer (.sub) and YouTube's preferred format - Scenarist Closed Caption (.scc). In 2009, a new feature that allows users to upload a simple text transcript of the spoken content and leave synchronization decision to the speech recognition mechanisms of YouTube was introduced. This feature saves users' time and trouble of transcribing the spoken video content and marking start/end times.

Figure 1 shows a YouTube video with automatic generated captions. These captions are superimposed on the video content, by clicking on the "CC button". As shown in the bottom part of the figure, close captions allow the video to be browsed by clicking on a given segment in the scrollable part of the entire caption track.



c++ code you can expect to be in the ballpark

Autoplay
Annotations
Speed: Normal
Subtitles/CC (2): English (auto-generated)
Quality: Auto 480p

Google I/O 2012 - Breaking the JavaScript Speed Limit with V8

Google Developers

Subscribe 1,019,436

121,718 views

Transcript

English (auto-generated)

8:08 on the v8 benchmark version 7 and you can see that it's gotten a lot better
8:13 since we launched chrome and we're now it at the position with v8 that we can
8:19 we can actually make a comparison against the performance of c and c++
8:24 code you can expect to be in the ballpark for certain code so don't give

Fig. 1. YouTube automatic captions

4 Proposed Methodology

Figure 2 presents the overall architecture of the implemented platform that includes the combination of a game-based annotation system and YouTube. After making YouTube videos available for the annotation application, and after having metadata contributed

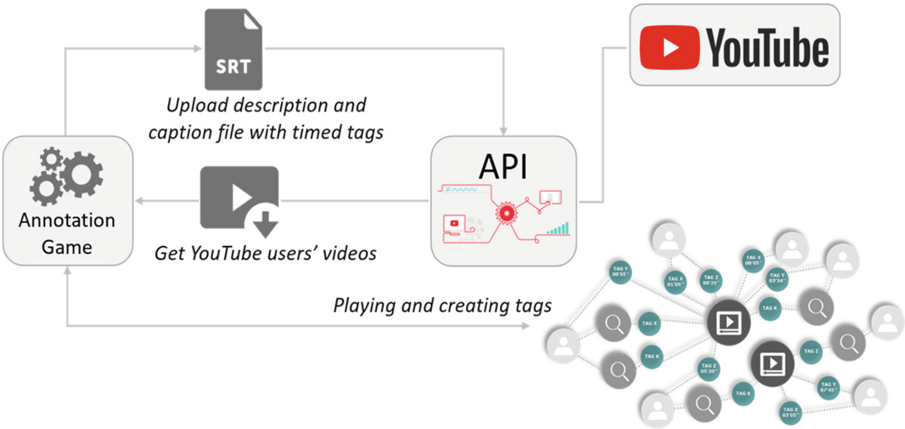

SRT
Upload description and caption file with timed tags
YouTube
API
Annotation Game
Get YouTube users' videos
Playing and creating tags

Fig. 2. Proposed system architecture

and validated by the crowd, this information is uploaded to YouTube in the form of captions and descriptions.

4.1 Annotation Game

The annotation system [12, 13] relies on a collaborative process and on gamification mechanisms to engage users on the tagging process. Tags may be freely introduced and players are rewarded if their contributions are considered valid. The created labels, or tags, are associated to specific time instants of the video, contributing to enhance the access to the exact moments of a video clip.

A scoring mechanism that takes into consideration past introduced information is used as an incentive for users to provide correct tags. Tag validation is achieved through a collaborative process, by analyzing the matchup between players' contributions. Additionally, semantically correlated tags are also considered, enabling enhancing and improving the quality of the dataset.

Three main aspects are considered on the process of tag validation: the tag itself and correlated tags from a dictionary; groups of tags organized in clusters; the number of times a tag, or correlated tag, appears in the respective cluster.

Clusters are groups of matching tags located nearby each other. They have a pre-defined length and are characterized by their centroid. Scoring is influenced by the distance of a tag to its centroid: 100, 50 and 10 points are considered. The higher the score, the closer the user is to the centroid. In this experiment, clusters were assigned windows of 12 s width, while scoring was linked to 8, 6 and 4 s time slots.

On the contribution of a player, the system verifies if the introduced tag is assigned to any of the existing clusters, or if a new one needs to be created. That assignment considers a pre-defined distance from the cluster centroid and its correlated tags. The introduced tag can result, or not, on the award of a score and on the validation of a tag if the number of agreements reaches a defined threshold.

To avoid player penalization for being the first one to introduce a specific tag, that later is validated by other players, an offline system is implemented to compensate this first effective contribution. An additional bonus is considered for having antedated useful metadata when the requirements for a score attribution are reached.

Besides contributing with metadata, players may also provide information that helps on the quality control of the tags. Moreover, different types of rewarding mechanisms that help on motivating good contributions and on maximizing the performance are considered in the game. This includes, besides scoring, prizes for completing an action, such as special badges, the definition of different game levels that the player may access and a leaderboard that shows his performance. Game levels are used on benefit of the annotation process as more difficult tasks (videos with less metadata are less likely to produce scoring) are provided to more qualified players.

The implemented mechanisms allow filtering erroneous information usually found on free tags and comments, to link metadata to timecodes and to have a collaborative effort on enhancing video information. A detailed description of the game functionalities is provided in [12, 13]. Aspects related to the performance, including usability, user engagement and tag accuracy, have been assessed in a user testbed [14].

4.2 Integrating the YouTube Platform and the Annotation Mechanism

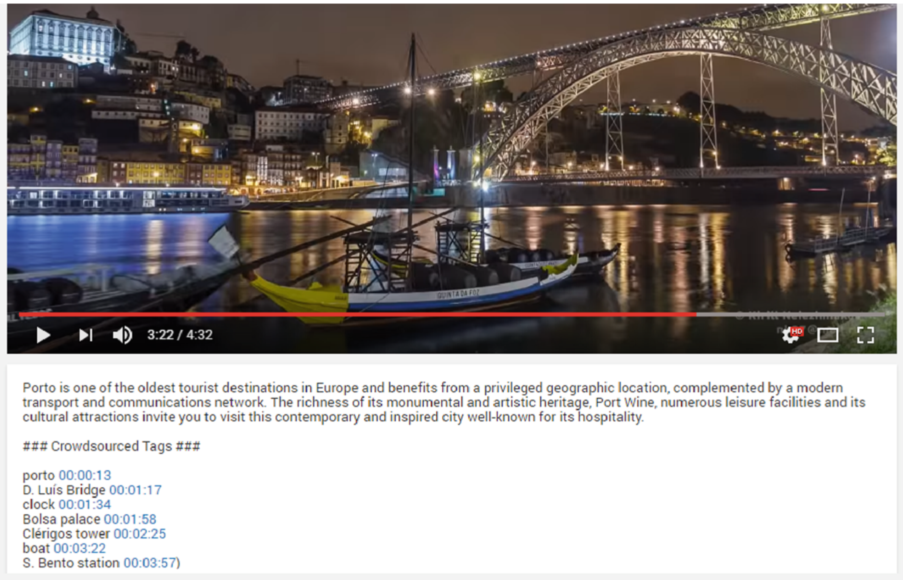
The YouTube Data API enables developers to incorporate a variety of YouTube functionalities into their applications. The API allows the communication with YouTube by providing the developer access to the videos and user information. This can be used to personalize a web site or application with the user's existing information.

As a first step and on user's consent, the system retrieves the videos on a YouTube account, integrating them in our platform for annotation. This is the only action a user willing to use the annotation game for enhancing his YouTube video description content is requested. Future requests on user's behalf are enabled by extracting an access token that is used for all the needed communication between our platform and YouTube, providing then a transparent, non-intrusive process. Token renewing is also automatic, enabling access to the user account even when he is offline.

On a first step, the video owner selects his YouTube videos for annotation. This process enables sharing those videos with the game engine by uploading the video IDs into the annotation system, making them available for the annotation community.

Inserted tags are initially stored locally to enable comparing and matching time-related information. When conditions for a tag to become validated are reached [12], the system will automatically create caption and description files that include the tag and the associated timecode, and will upload this information to the YouTube account making this metadata fully available for everyone.

Figure 3 illustrates the use of the description field in YouTube, enhanced with timed tags imported from the annotation game. The initial description provided by the owner is kept but the field is updated with the navigable tags that, besides helping enhancing



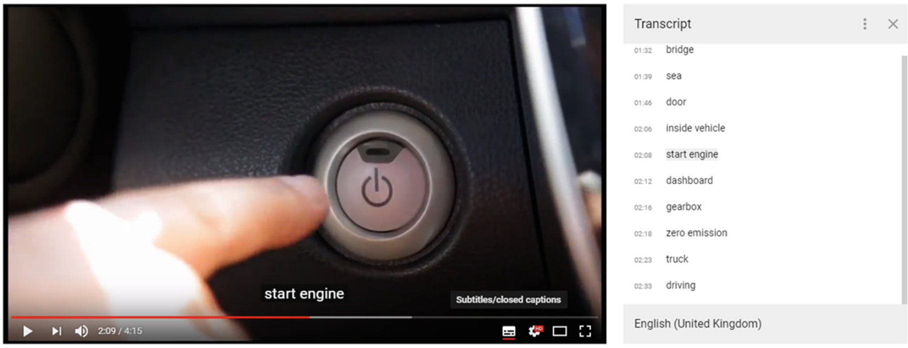
Porto is one of the oldest tourist destinations in Europe and benefits from a privileged geographic location, complemented by a modern transport and communications network. The richness of its monumental and artistic heritage, Port Wine, numerous leisure facilities and its cultural attractions invite you to visit this contemporary and inspired city well-known for its hospitality.

Crowdsourced Tags

porto 00:00:13
D. Luís Bridge 00:01:17
clock 00:01:34
Bolsa palace 00:01:58
Clérigos tower 00:02:25
boat 00:03:22
S. Bento station 00:03:57)

Fig. 3. Crowdsourced timed tags description

search precision, enable jumping to exact moments on the video. Figure 4 presents the use of captions to increment metadata. Tags may be overlaid in the video presentation and, additionally, they are listed to enable hyperlinking to video instants.



01:32	bridge
01:39	sea
01:46	door
02:06	inside vehicle
02:08	start engine
02:12	dashboard
02:16	gearbox
02:18	zero emission
02:23	truck
02:33	driving

Fig. 4. Timed caption track

4.3 System Work Flow Process

The data flow process is depicted in Fig. 5 showing the interactions between the annotation system, the browser and the YouTube API. The process can be summarized as:

- User logs into our platform and chooses the built-in functionality for sharing his YouTube videos with our system.
- To enable retrieving his videos from the YouTube, and later on to publish captions and descriptions under his account, an authorization is required. This will allow our system to use the YouTube API methods. The user's Google Account is used to authorize the application to access the videos and to upload metadata.
- The OAuth2 authorization process is initiated on the first attempt to use the functionality in the game. On user consent, Google returns an access or/and a refresh token that is/are stored in the database for future use. This token allows uploading information from the annotation process into the user YouTube account.
- User' videos are listed and can be selected and shared with the annotation platform.
- Video IDs are extracted and added to our database to make them available for the crowd and playable along with other existing videos.
- According to the user gaming level, videos from our database will be retrieved in order to be presented to the user and played during the game.
- During the game and following the validation rules, tags can become valid due to exact or concept matching.
- A YouTube compatible.srt (SubRip) caption file and a description, both including timed tags, are automatically generated. The original description is merged with the new description so that no information is lost.
- Data is uploaded to the YouTube account by using the user ID and the token.
- Caption track and description become publicly available and ready to use/view.

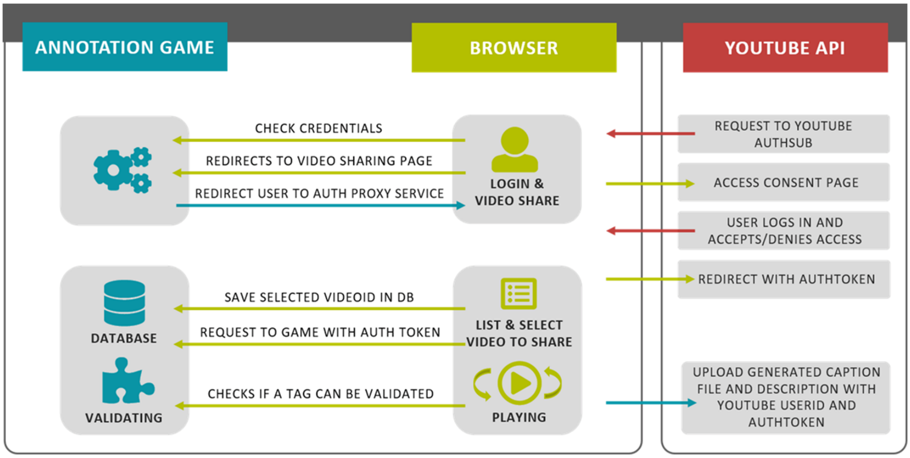

ANNOTATION GAME
BROWSER
YOUTUBE API
CHECK CREDENTIALS
REDIRECTS TO VIDEO SHARING PAGE
REDIRECT USER TO AUTH PROXY SERVICE
LOGIN & VIDEO SHARE
REQUEST TO YOUTUBE AUTHSUB
ACCESS CONSENT PAGE
USER LOGS IN AND ACCEPTS/DENIES ACCESS
REDIRECT WITH AUTHTOKEN
DATABASE
SAVE SELECTED VIDEOID IN DB
REQUEST TO GAME WITH AUTH TOKEN
LIST & SELECT VIDEO TO SHARE
VALIDATING
CHECKS IF A TAG CAN BE VALIDATED
PLAYING
UPLOAD GENERATED CAPTION FILE AND DESCRIPTION WITH YOUTUBE USERID AND AUTHTOKEN

Fig. 5. Data workflow

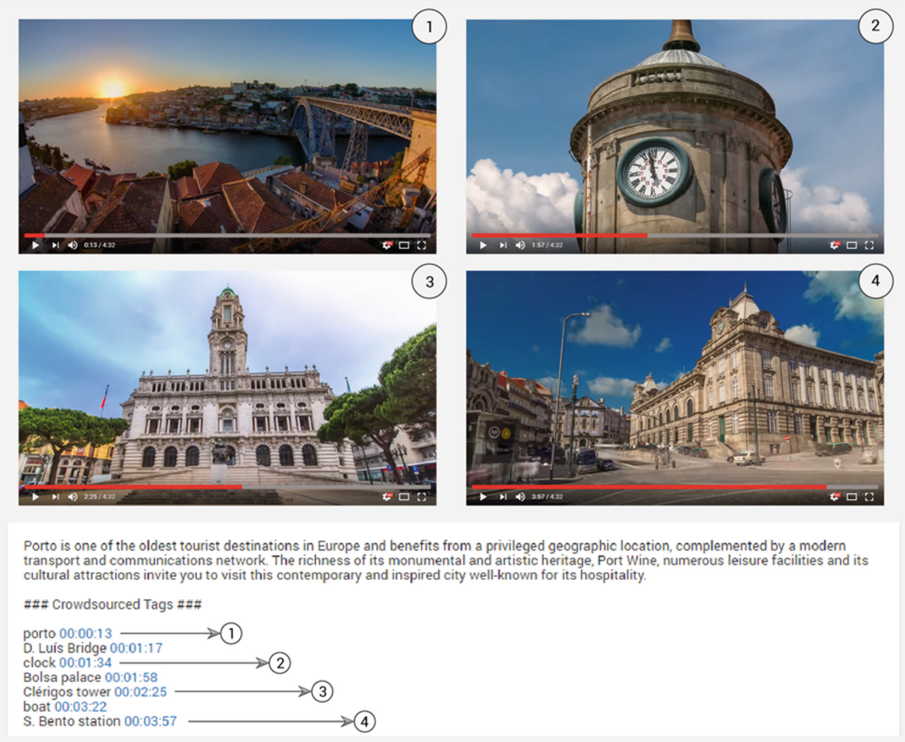
5 System Evaluation

We have conducted a quantitative and qualitative experiment to evaluate the performance of the system. To assess the annotation approach, a set of volunteers that simulated a crowdsourcing environment was asked to interact with the system. Different genres of content were considered to make the annotation process not focusing in just one kind of material. Besides analyzing the quality of the annotations, this experiment enabled also checking the usability of the system and collecting feedback from the users by means of a survey.

Findings show that players tend to be very accurate in time when typing some tag: 92.4% of the users were very consistent and introduced the same tag, or correlated tags, near other players' tags. This can be explained by the fact that the gamification approach encourages users to provide accurate information in order to score and progress in the game. 60% of the contributed tags were validated by the system showing tag matchup. These contributions allowed indexing 71 moments of the videos.

Motivation and engagement features included in the annotation process proved to be effective. Not only accuracy was achieved, as well as it was evident that players interacted actively with the game, competition within the top positions was acknowledged and answers to the questionnaire enable identifying motivation and enthusiasm (82% of the players declared having enjoyed the game and it's features) [14]. These findings are quite important, as productivity will depend on the motivation and enthusiasm that the gamification concepts can provide.

The metadata obtained collaboratively is expected to make searching and navigation in video archives more efficient and to reduce the need for professional and expensive processes of describing content. Figure 6 presents the navigation efficiency increase achieved by uploading into YouTube the crowd contributed tags resulting from the experiment. Direct access to specific parts of the same video was enabled by navigating on hyperlinked tags introduced in the description field.



Porto is one of the oldest tourist destinations in Europe and benefits from a privileged geographic location, complemented by a modern transport and communications network. The richness of its monumental and artistic heritage, Port Wine, numerous leisure facilities and its cultural attractions invite you to visit this contemporary and inspired city well-known for its hospitality.

Crowdsourced Tags

porto 00:00:13
D. Luís Bridge 00:01:17
clock 00:01:34
Bolsa palace 00:01:58
Clérigos tower 00:02:25
boat 00:03:22
S. Bento station 00:03:57

Fig. 6. Example of a YouTube video with timed tags obtained collaboratively

6 Conclusion

This paper proposes the use of a collaborative annotation game to collect metadata contributed by the crowd and enhance content in the YouTube platform using the caption feature and the description field of YouTube. By using a validation mechanism that enables cleaning erroneous information before uploading it to YouTube, the approach has benefits over the popular use of comments. To the best of our knowledge this is the first implementation of a system that uses YouTube caption and description features to improve YouTube search results using a collaborative approach. Apart from adding keywords to the video, it also associates timecodes to video descriptions, enhancing the navigation on YouTube content. Our method tries to solve two problems: the lack of useful metadata for accurate video retrieval and the difficulty on video navigation due to the lack of timed descriptions. Future work includes adding others video sharing platforms besides YouTube and creating a browser extension that allows the player to directly play the game on YouTube without the need to access another application.

Acknowledgements. The work presented was partially supported by FourEyes, a Research Line within project “TEC4Growth: Pervasive Intelligence, Enhancers and Proofs of Concept with Industrial Impact/NORTE-01- 0145-FEDER-000020” financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

References

1. Zhou, R., Khemmarat, S., Gao, L., Wan, J., Zhang, J.: How YouTube videos are discovered and its impact on video views. Multimedia Tools Appl. **75**, 6035–6058 (2016)
2. Guimarães, R., Cesar, P., Bulterman, D.C.A.: Let me comment on your video: supporting personalized end-user comments within third-party online videos. In: Proceedings of the 18th Brazilian Symposium on Multimedia and the Web, pp. 253–260. ACM, Brazil (2012)
3. Yamamoto, D., Masuda, T., Ohira, S., Nagao, K.: Video scene annotation based on web social activities. IEEE Multimedia **15**, 22–32 (2008)
4. Mulholland, E., Kevitt, P.M., Lunney, T., Schneider, K.-M.: Analysing emotional sentiment in people’s YouTube channel comments. In: Interactivity, Game Creation, Design, Learning, and Innovation, pp. 181–188. Springer, Cham (2016)
5. Yamamoto, D., Masuda, T., Ohira, S., Nagao, K.: Collaborative video scene annotation based on tag cloud. In: Huang, Y.-M.R., et al. (eds.) Advances in Multimedia Information Processing - PCM 2008, pp. 397–406. Springer, Heidelberg (2008)
6. Khusro, S., Khan, M., Ullah, I.: Collaborative video annotation based on ontological themes, temporal duration and pointing regions. In: Proceedings of the 10th International Conference on Informatics and Systems, pp. 121–126. ACM, Giza (2016)
7. Gatteschi, V., Lamberti, F., Sanna, A., Demartini, C.: An audio and image-based on-demand content annotation framework for augmenting the video viewing experience on mobile devices. In: Proceedings of 2015 IEEE International Conference on Mobile Services, pp. 468–472 (2015)
8. Steiner, T., Verborgh, R., Van de Walle, R., Hausenblas, M., Vallé: crowdsourcing event detection in YouTube video. In: Proceedings of the 1st workshop on detection, representation, and exploitation of events in the semantic web (2011)
9. 3Play Media: Adding Closed Captions to YouTube Videos Increases Views. http://www.3playmedia.com/customers/case-studies/discovery-digital-networks
10. Captions for YouTube. http://www.captionsforyoutube.com/
11. Burgess, J., Green, J.: YouTube: Online Video and Participatory Culture. Wiley, Hoboken (2013)
12. Pinto, J.P., Viana, P.: TAG4VD: a game for collaborative video annotation. In: 2013 ACM International Workshop on Immersive Media Experiences, pp. 25–28. ACM, Spain (2013)
13. Pinto, J.P., Viana, P.: Using the crowd to boost video annotation processes: a game based approach. In: Proceedings of the 12th European Conference on Visual Media Production, p. 22:1. ACM, London (2015)
14. Viana, P., Pinto, J.P.: A collaborative approach for semantic time-based video annotation using gamification. Hum. Centric Comput. Inf. Sci. **7**, 13 (2017)