Viana and Pinto *Hum. Cent. Comput. Inf. Sci.* (2017) 7:13
DOI 10.1186/s13673-017-0094-5

Human-centric Computing and Information Sciences

RESEARCH Open Access

A collaborative approach for semantic time-based video annotation using gamification

Paula Viana[1,2*] and José Pedro Pinto[1]

*Correspondence:
paula.viana@inesctec.pt
[1] INESC TEC, Campus da FEUP, Rua Dr. Roberto Frias, 378, 4200-465 Porto, Portugal
Full list of author information is available at the end of the article

Abstract

Efficient access to large scale video assets, may it be our life memories in our hard drive or a broadcaster archive which the company is eager to sell, requires content to be conveniently annotated. Manually annotating video content is, however, an intellectually expensive and time-consuming process. In this paper we argue that crowdsourcing, an approach that relies on a remote task force to perform activities that are costly or time-consuming using traditional methods, is a suitable alternative and we describe a solution based on gamification mechanisms for collaboratively collecting timed metadata. Tags introduced by registered players are validated based on a collaborative scoring mechanism that excludes erratic annotations. Voting mechanisms, enabling users to approve or refuse existing tags, provide an extra guarantee on the quality of the annotations. The sense of community is also created as users may watch the crowd's favourite moments of the video provided by a summarization functionality. The system was tested with a pool of volunteers in order to evaluate the quality of the contributions. The results suggest that crowdsourced annotation can describe objects, persons, places, etc. correctly, as well as be very accurate in time.

Keywords: Tagging, Video annotation, Gamification, Crowdsource, Metadata

Background

Tagging systems enable new modalities of social communication and opportunities for data mining [1]. By being engaged in the annotation process, humans contribute to index information, and it is likely that they are attracted to retrieve information as well. However, a great part of that user-generated annotations have no quality-control process that guarantees the effectiveness of the data collected. Tags created by single users are typically noisy and selfish, contain misspelled words, miss important keywords and are not linked to specific timecodes [1–3]. This limits the usefulness of the tags, especially for efficient access to large assets of video content, and, more concretely, it does not allow the accurate access to exact parts of a video.

Techniques grounded on video processing are still mostly not feasible given the computer resources required and the difficulty to develop a reliable and universal system to any type of content. Nevertheless, random access to crucial points of videos is essential for retrieving the best content and data that fills our expectations.

Many TV and radio broadcasters accumulate very large archives, with various degrees of granularity of content and available metadata. Some of them are very old, speechless

and do not have metadata associated. Most of cataloguing efforts have been geared towards reuse, to enable programme makers to easily find snippets of content to include in their own programmes. However, creating this metadata is a timely and expensive process, still relying on experts that continuously verify and sort the contents to be published, providing structured information for describing the content as a whole, as well as specific descriptions of parts of the video.

Some broadcasters are already implementing tagging mechanisms using e.g. *DBpedia* to help audience to easily find programmes [4]. However, the current tools are mostly used to suggest tags to editors, based on the metadata already known (title, description, etc.), not providing effectively new information.

Crowdsourcing has been gaining points as a method to collect metadata descriptors, adding extra alternative textual information to the one that already exists. The idea of collecting metadata using gamification concepts has been applied on some video footage and has attracted players to entertainingly contribute to the cause [5–7].

This paper presents a platform for video content annotation using a collaborative approach. Based on the concepts of crowdsourcing, gamification and participatory media, the system collects metadata that enables describing a video asset (Fig. 1). Information is linked to specific time instants contributing to enhance search and access to video content. Accuracy of the tags is achieved through a validation and scoring mechanism that awards players based on their success and implements a motivation scheme. As a motivational complement, users are also encouraged to reach pre-define goals that enable them to unlock new badges and progress in levels of difficulty.

New features, that include better algorithms for tag validation, new scoring and motivation mechanisms, a dictionary that helps achieving better descriptions and enhanced navigation features have been added to previously published work described in [8, 9]. The approach was validated through a user experiment and findings suggest that not only the users feel engaged and willing to contribute, but also that collected data is valid and correctly describing key concepts of the video. The proposed solution is expected to

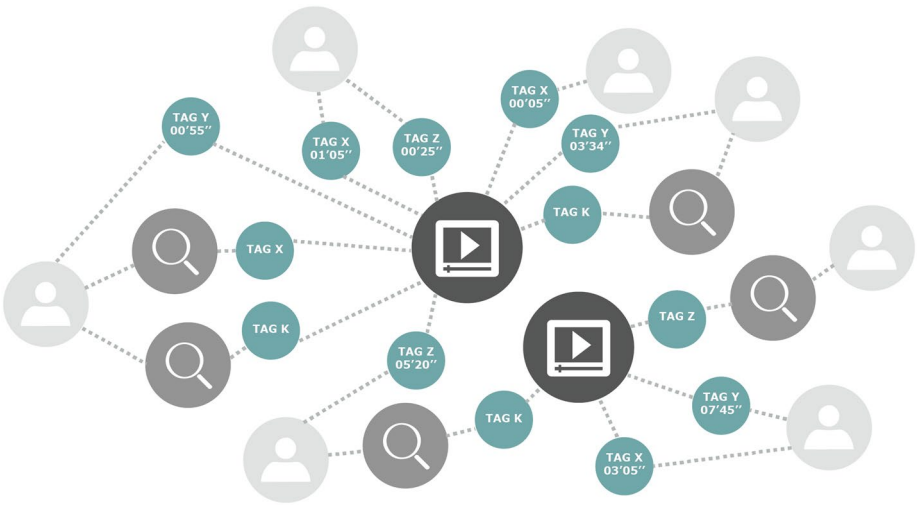

Fig. 1 Crowdsourcing video timed tags

have a significant impact on the management of valuable video assets that would otherwise be unavailable for use.

The rest of the paper is organised as follows. In the next section we present the state of the art approaches for video annotation focusing on collaborative and content analysis methodologies. In "Game explained" we describe our proposal, detailing the main game concepts, motivation mechanisms implemented to assure active participation of the crowd and that include, among others, different scoring and game levels, a leaderboard and a set of awards for the accomplishment of tasks. The mechanisms for tag validation are also described. The evaluation testbed and results are presented and discussed in "System evaluation". Finally, we highlight some conclusions and directions for future work.

Related work

Searching and browsing large collections of video assets depends greatly on the capacity of describing this content. Several approaches have been proposed in the literature to enhance the accuracy of search queries. In this chapter, we will focus on two main methodologies: the implementation of collaborative based mechanisms to collect metadata and content analysis approaches to extract relevant information from the media.

Collaborative approaches for content annotation

Game based approaches

Games with a purpose (GWAPs) are an example of an emerging class of games that uses gamification and human computation power to collect data from the interaction with human users. This idea was firstly introduced in the domain of multimedia content by Luis von Ahn with the *ESP Game*. This multi-player game harnesses human abilities to label images and, based on the consensus among users, it provides a method to ensure the quality and consistency of the labels. The idea to collect metadata through games has been applied to video, audio, and images archives [7, 10–14].

ESP Game has served as a prototype for many later successors developed by Galleries, Libraries, Archives and Museums. Brooklyn Museum's online game *Tag! You're it!* [14] is a crowdsourcing game based on the *ESP* platform. Aligning fun with the need to help tagging objects to better search their collection, *Digitalkoot* [10] engages users to correct the optical character recognition (OCR) output of scanned documents. Users are asked to validate whether the digital text corresponds to the image of the word. The more words a user validates, the higher score gets. In *Stupid Robot,* players score points by teaching a robot about what they see in displayed images. It contributes with data to libraries and museums' digital collections and makes images more accessible for everyone [6, 15].

Examples where users are invited to contribute with metadata to describe video content can be found in different areas of applications. *CrowdSport* [12] is a system that makes use of wisdom of the crowd to annotate video content of soccer games. The users are paid to annotate short video clips with semantic events: spot elements of the team, the time, the position on the field, etc. To ensure the quality of the user's data, a user is first rated by his peers and, since each sequence is annotated multiple times by different users, metadata is integrated and compared among contributors and rating is adjusted

accordingly. *Guess What?* [11] is a *Facebook* app game where players are invited to watch a video clip and then to answer to questions about it. The main objective is to receive points, taking into account other users' guesses. On *NexTag,* players are presented a random short video clip and engaged to list what's being shown on the media. Score is awarded according to matched tags or new tags introduced [6]. *Waisda?* [5] is possibly one of the most complete and exciting games in the area of content annotation, and aims to tag what you see or hear, regarding some video content. The basic scoring mechanism is tag agreement, with two players entering the same tag within pre-defined timecode distances from each other. The matching process has been enhanced by importing pairs of words from tag similarity lists and dictionaries for specific typed tags.

Attracting and holding enough participants is a critical issue to assure the effectiveness of game based approaches. Systems with just a few players are not suitable to generate meaningful annotations and, consequently, to generate contextual and representational quality data. The implementation of motivation mechanisms is then an important feature that contributes to increase the effectiveness of the game [7, 16].

Other collaborative approaches

Crowdsourcing approaches not using gamification concepts can also be found in literature. Davis et al. [17] presents a media tagging system which integrates social networking with online media. The user interaction metadata is garnered and aggregated to form semantic metadata to a given video.

The multimedia search engine from [18] facilitates semantic access to rock 'n' roll concert videos. By using crowdsource techniques with a combination of automated content understanding and the wisdom of the crowds, they have shown how beneficial crowdsourcing can be to a video search engine that automatically recognizes video fragments on a semantic level.

Automatically suggesting tags from web blogs has been proposed by Mishne [19] and Sood et al. [20]. The system works by finding similar posts, filtering and re-ranking results based on tag co-occurrence and frequency, improving the suggestions on tagged posts. Wu et al. [21] follows a similar concept by presenting an application that extracts time-sync video tags by exploiting crowdsourced comments from video websites.

User search behaviour has also been used as a process to enrich video assets. Yao et al. [22] explores the click-through data for learning video relationships and semantic video similarity and, consequently, enrich video tags.

Although the contribution of timed tags to increase the efficiency of media content access is universally accepted, some authors provide concrete analysis of the impact of those tags. The video data set provided in Xu and Larson [23] contains social videos and user-contributed timed tags. Using the *Viddler* platform for the extraction of timed tags, their results confirm the importance of timed metadata. Based on deep-link comments extracted from *YouTube*, [23] analyses how the viewers deep-link and how it opens up the possibility of a new relevance criterion for non-linear video access.

Content analysis approaches

Proposals based on multimedia content processing have also been exploited. Examples based on image, video, audio or text analysis can be found applied to several areas of

application. Bertini et al. [24] uses ontologies and visual information to automatic annotate soccer highlights. Moxley et al. [25] exploits the overlap in content of video news, to automatically annotate and improve original annotations. Larson et al. [26] uses three different tasks to automatically match episodes with labels from a keyword thesaurus, predict tags that are assigned by users to their online video and assign geo-coordinates to videos.

Tagging video sequences requires a lot more processing power than tagging single images. Some researchers have proposed automatic, or semi-automatic processes, that based on relevant frames [27–29] and similarity between frames, as well as relationships between them [30, 31], can suggest and add new tags to be associated to videos on a key-frame/scene level.

ShotTagger [32] uses a combination of context and content based methods to annotate the shots corresponding to the same tag within an Internet video. Based on co-occurring tags and temporal smoothing, it refines the annotations enabling consistency across shots. The results have showed the feasibility and effectiveness of tag annotation and tag-based shots, as well as, how location tags into video can be built into a tag-based video browsing.

Based on collective knowledge and visual similarity of frames, [30] presents a system for video tag suggestion and temporal localization. The developed algorithm suggests new tags that can be associated to video content, based on the visual similarity of frames extracted from social websites like *YouTube* and *Flickr*. Other systems as [31, 33] use similar approaches based on social knowledge, and combine visual similarity, tag frequency and geo-localization to suggest or create new and relevant tags to enrich the annotation quality.

Crowdsourced data from social multimedia applications hosts thousands of diversified semantic tags, which allow many systems to rely on tag frequency and semantic correlation to create unsupervised annotations. Tran et al. [34] corrects and complements users tags, comparing directly the visual content of the videos, using different sets of features such as Bag-of-visual-Words or frequent patterns. Tags are then propagated between visually similar videos, according to the frequency of these tags. Altadmri and Ahmed [27] matches events or objects identified in video clips with similar content available in a dataset by using low-level visual descriptors. Pre-annotated metadata available in the dataset is then used as input to a dictionary and a semantic concept mapping algorithm that provides final content annotation is implemented. Nga and Yanai [29] integrates visual information of video shots and tag information from Web videos in order to automatically extract relevant video shots of a specific action.

Text recognition or OCR has also been explored by several researchers [35, 36] to detect and recognize text on video images to improve browse and search in a video asset.

Automatic speech recognition systems have addressed the tasks of automatically generating labels that characterize the content of spoken multimedia files. Moxley et al. [25], Larson et al. [26], Yang and Meinel [37] use automatic speech recognition from audio tracks to enhance multimedia content metadata and facilitate search.

The work described in Eggink and Raimond [4] provides a solution that implements an automated semantic annotation that classifies mood, sound effect, and semantic tagging,

targeted to the general public. This solution is supported by sound effects based on subtitles and speech recognition software.

Overview

The main drawbacks of these proposals are either related to computation costs associated to multimedia processing or to the poor mechanisms implemented to guarantee the quality of the tags which are usually too generalist, not associated to a timecode and not correctly describing the content. Additionally, when relying on collaborative scenarios, motivation mechanisms that contribute to increase the number of contributions are not considered.

The work presented in this paper proposes a solution for video content annotation based on a collaborative process that uses the concepts of crowdsourcing and gamification to collect metadata. A complex scoring and validation systems that includes the functionality of collaboratively rejecting previously introduced tags and using aggregated information from group of players, enables enhancing the quality of the metadata accepted by the system. Information is linked to specific time stamps contributing to enhance search and access to video content. Previously proposed solutions based on the concept of games with a purpose tried to bridge the semantic gap between descriptions by using dictionaries or similarity lists. However, inputs requested to the player are, in some cases, still too complex (not just individual tags but formal sentences are required) and semantic concepts are only identified through standard dictionaries and thesaurus. Our approach goes beyond these solutions and uses common crowdsource methodologies for the creation of new metadata along with crowdsource tag-based dictionaries. Moreover, and thanks to the web-based and HTML5 (Hypertext Markup Language revision 5) technologies, the community of contributors can easily participate using different devices, instead of using proprietary players with much slower performance and portability.

Game explained

On *Tag4VD* (Tag for video) users can interact with videos in two different modes: the game competition mode or by searching and browsing the existing content. Both modes are available for registered and guest users but some functionalities are not included in the guest mode (score storage, rewards unlock, new videos and levels, etc.). None of the information provided by the guest users is, however, discarded and can help registered players along their games.

In the game competition mode, players are presented a set of random video clips, selected according to their game level, which they are required to annotate. While playing, users get updated information on the remaining time and about their performance, presented as scoring and as a list of valid tags inserted in the current session (Fig. 2).

Besides contributing with metadata, players may also provide information that helps on the quality control of the tags introduced. In this other mode, users can navigate through previously annotated video and, for shown tags, provide their option on how that concept really describes content (Fig. 3). They can also signalize parts of the video as containing interesting/important moments.

Fig. 2 Game competition mode

Fundamental competition mode mechanisms

Tag4VD is a single player game that asks players to label some random videos within a pre-defined game duration time. The created labels, or tags, are anchored to particular time stamps of the video, contributing to enhance the access to the exact instant of a video clip when executing a query or browsing the dataset. Tag validation is achieved through a collaborative process, by analyzing the matchup between players' contributions in the same or nearby clip instants [8]. For each validated tag, the player is rewarded for his contribution. To make the game fairer, the system keeps track of all the tags introduced during the game, and scoring can be assigned either while a player is still actively playing or on offline mode, if new information confirms previously introduced tags.

Figure 4 illustrates the process associated to a tag submission on *Tag4VD*. The tag, together with auxiliary information that includes the timecode, identification of the video, of the player and of the game, is used to query the database for same tags associated with that video and nearest cluster. The concept of clusters is used to aggregate identical tags within a pre-defined timeframe. A dictionary is also used to enable identifying correlated concepts that are used as synonymous on the scoring and validation process.

Three main aspects are considered on the process of tag validation: the tag itself and correlated tags from the dictionary; the existing clusters and the size of the matching cluster.

Fig. 3 Game crowd judgment mode

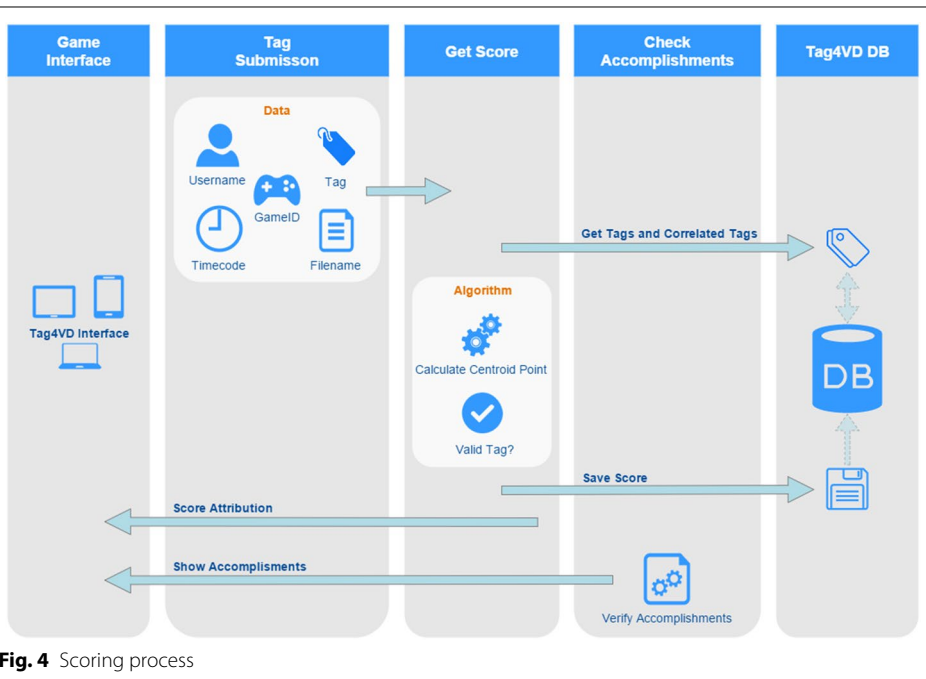

Fig. 4 Scoring process

A non-zero scoring is saved in the database and displayed on the game interface if a tag match is found within a pre-defined timeframe. Besides this scoring award, accomplishment of tasks or other objectives are checked and additional awards in the form of badges or game levels may be given.

These aspects are discussed in detail in the next section.

Scoring

Tag4VD uses a scoring mechanism to motivate the players based on how close in time a reasonable number of players introduce the same semantically related tag to describe a scene. Tag agreement is based on the exact match of tags or on the match of semantically similar tags available in our dictionary. This allows the system to extend the default syntactic tag matching with semantic similarity matching.

Clusters are used to group tags that, although not associated to the same exact timecode, are located nearby each other. Given that the same tag can co-occur in different instants of the video clip, but describing a different circumstance, the system keeps track not only of the tags associated to a video stream but as well of their timecodes. Clusters are defined by its centroid (the mean value of all the timecodes of the tags belonging to that cluster) and by the tag-id that identifies the concept being described by that cluster. Multiple clusters associated to the same concept may exist throughout the video.

Whenever a tag is introduced, the system verifies if that tag can be assigned to any of the existing clusters or if a new cluster needs to be created. Associating a tag to a cluster requires tag syntactic or correlation matching with the tag-id of a cluster whose centroid is within a pre-defined distance of 6 s from the tag been processed. When a cluster's population is modified, a new centroid is computed.

The centroid of a cluster is also a crucial piece for the attribution, or not, of a score. For a player to be rewarded for his tag, the following requirements must be fulfilled: that tag must have been assigned to a cluster based on the previously described process; this cluster must be larger than three elements; and the tag under evaluation must be less than 4 s from the cluster's centroid. In such case, the player will be awarded one of three score levels: 100 points for each tag inserted within a distance of 2 s from its cluster's centroid; 50 points for tags within a range of 3 s of the centroid; and 10 points when the distance to the centroid is up to 4 s [8].

Figure 5 illustrates this process of cluster and score attribution. In the example, 3 clusters can be identified: "Cluster X-1" and "Cluster X-2" that although describing the same concept "X" are located in different instants of the video; "Cluster Y-1" describing another concept. Each of these clusters has a number of tags no more than 6 s from the

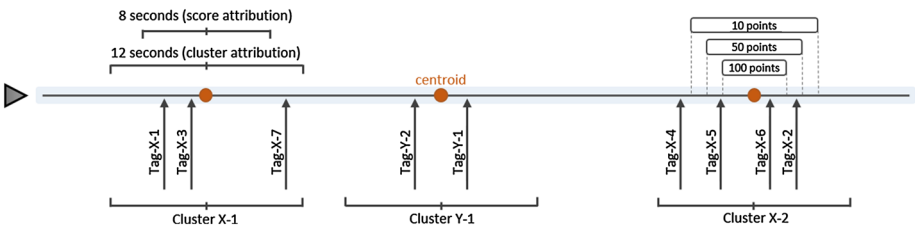

Fig. 5 Score and cluster assignment

centroid. When a tag is introduced by a player, the system searches for the closest cluster matching that concept, trying to assign that tag to an existing cluster. Taking "Tag-X-7" as an example, the system identifies "Cluster X-1" (already containing "Tag-X-1" and "Tag-X-3") as the candidate to incorporate this new tag. No score is however processed as the initial cluster had only 2 tags. In the example depicted, only "Cluster X-2" enabled rewarding players. Based on the distance of each contribution to the centroid, "Tag-X-6" will be awarded 100 points and both "Tag-X-2" and "Tag-X-5" will receive 50 points. "Tag-X-4", although associated to that same cluster, will not get any points since it is out of the specified range.

To avoid player's penalization for being the first one to introduce a specific tag, that later is validated and scored by other players, an offline system is implemented to compensate the firsts effective contributors—when the conditions for score attribution are reached, an additional bonus (200 points) is even considered for the players who had antedated useful metadata. For the example in Fig. 5, if a new tag is inserted in "Cluster-X-1", tags "Tag-X-1", "Tag-X-3" and "Tag-X-7" will become valid and those players will receive the 200 points bonus. Given that the centroid cluster may also change, due to new users' tags on different positions, a background tracking mechanism is responsible to automatically consider these situations and update players' scores.

Explicit crowd judgment mode

In this game mode, crowd's opinion is used as an additional mechanism for validating metadata introduced in the competition phase. While browsing the asset, players may provide their opinion on the quality of existing tags through a simple "like/dislike" judgment. This additional information is used to discard wrong tags and make stored information more accurate.

For this process, users are assisted by several functionalities:

- A tag cloud that enables a first insight on the frequency of tags, while providing a mechanism to direct the user to specific points related to the clicked tag. Furthermore, and based on the number of validated clusters for the clicked tag, the system can also create multiple points of interest (Fig. 6) that help the user to rapidly access them.
- A summarization feature (Fig. 7) that can create a shorter version of the video with the top ten interesting moments based on the likes bar chart created by the community of players.
- A like bar chart, where each bar provides information on the time instant, and that can identify the most impressive moments. By pressing on one of those bars, users are routed automatically to that part of the video.

Motivation mechanisms

Besides scoring, other motivation mechanisms were implemented to motivate participation.

Badges can be used to compensate accomplishments, and keep player engaged, or to direct users to perform some task that can be significant for the annotation process or even for providing feedback to the game designer.

Fig. 6 Navigation features

Fig. 7 Video summarization based on key points of the video

Difficulty levels reflect users' expertise and introduce more challenging tasks, making the process of scoring more difficult. By showing a progress indicator it can also trigger competition to reach a higher position on the leaderboard or to have recognized some prestige.

On *Tag4VD* different motivations are implemented (Fig. 8) in order to promote fun, participation, achievements, persistence and enjoyment:

- Rewards for specific actions which may have impact on the implemented systems, game and information retrieval. Table 1 shows the available badges to unlock in *Tag4VD*.

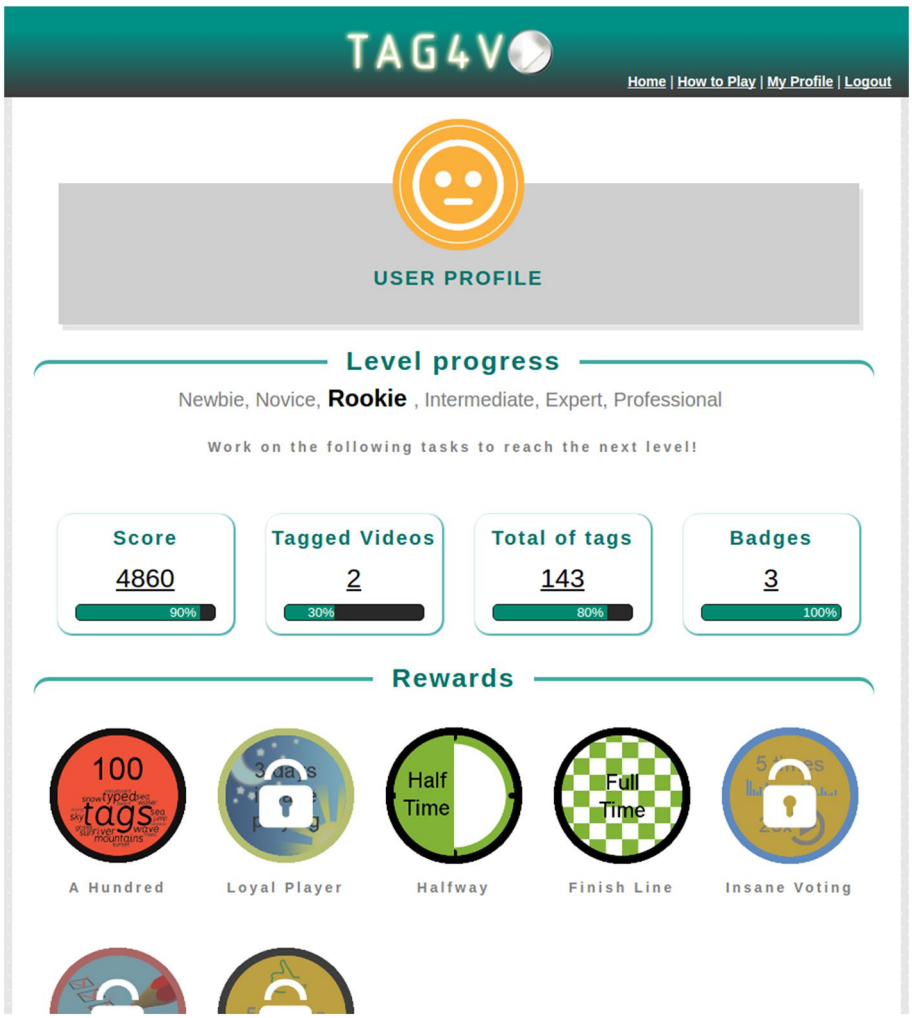

Fig. 8 Player Profile

- Levels of difficulty based on total score, number of tagged videos, number of tags, and badges unlocked.
- Videos to annotate are grouped in different classes of difficulty: players on a lower level of the game will be presented videos with more annotated content, so they can have more chances to score and progress on the game.
- Tips are used to help players achieving better results.

To enable collecting information that may help improving the game, players are invited to fill in a questionnaire and provide their gaming experience feedback, including positive affect, immersion and challenge. The survey contains questions about how they felt about using the system, how enjoyable it was, how effective the motivation factors are, suggestions on improvements, etc. As a reward for the user, the system will unlock the "Fill form" badge (Table 1) and consequently help him reaching a higher level.

Table 1 Badges examples

Badge name	Badge description
100 typed tags	Reach 100 tags entries
3 days insane playing	Played 3 days consecutively
Half time	Played 1/2 of the game total duration
Full time	Played until game is over
5 times voting in 20 videos	Vote 5 times on interesting moments of at least 20 videos
Fill the form	Fill our form and submit it
Like and dislike >50 times	Agree or disagree more than 50 times on popup tags

Tag correlation

Trying to match words' meaning and concepts has already been considered in some applications related to content annotation. *Waisda?* [5] based its score attribution on matching tags between players but also on tag similarity lists and dictionaries, which include synonyms, specific tags or specific types of tags according to the video category. However, those lists and dictionaries have to contain specific metadata to categorize it to a video domain.

WordNet or DBpedia are open vocabulary based dictionaries that may be used to find correlation between words. However, they mainly intend to provide a synonymous platform and other concepts that can be correlated and be used to describe content are not inferred from these systems. Attempts to use multimedia domain lists of words and dictionaries led to the creation of specific lists such as NUS-WIDE [38] a popular web image dataset extracted from Flickr and that includes approximately 260k images with a manual annotation of 81 concept categories.

For implementing the tag correlation mechanism mentioned before, we used the NUS-WIDE database and the *Web 2.0 Dictionary* approach [39]. The main idea is to build a dictionary able to provide correlation ranges between words that can be used to associate different concepts. As an example, when inserting the tag "car" in the competition mode, the system will promptly correlate it to equivalent words like "auto", "cars", "automobile", "vehicle", etc. identifying not only singular/plural relations but also similar words used in social media applications. An even more illustrative example could be on the co-relation that can be obtained from this dataset concerning e.g. "beach", "lifeguard" and "sun umbrella". Although none of these words are synonymous, they describe concepts that frequently occur together.

Considering the methodology described by Yang et al. [39] we have created our dictionary based on the metadata from NUS-WIDE database. Annotations from the multimedia content are organized in a group of "bags", each bag relating to a specific multimedia file (in our case Flickr images are used). For each unique parent tag, we have considered all bags of tags that contain that same parent tag. The correlation between all the child tags, other than the parent tag, and the parent tag, is calculated by the frequency of occurrence of each of the child words appearing in the bags of tags. The list of child words is sorted by their co-occurring frequency. To reduce noise, we have decided to limit the minimum correlation value to 0.1. This means that the correlation dictionary, stored in our database, has correlation values in the range of 0.1–1.

Table 2 Correlation tags examples

Dog		Car		Cinema		Airplane	
Tag	**Correlation**	**Tag**	**Correlation**	**Tag**	**Correlation**	**Tag**	**Correlation**
Dogs	0.279470	Auto	0.184225	Movie	0.320916	Aircraft	0.553820
Puppy	0.239072	Cars	0.182570	Film	0.257879	Plane	0.481101
Pet	0.181125	Automobile	0.153612	Actor	0.174785	Aviation	0.414543
Animal	0.135099	Vehicle	0.104522	Display	0.126074	Flying	0.341413
Puppies	0.118874	Street	0.103971	Theatre	0.123209	Airport	0.335661
Pets	0.106291	Racing	0.101765	Theater	0.114613	Jet	0.308134

We have considered a total of 269,648 Flickr images with 424,853 relation tags. Table 2 presents some tag correlations examples and their value of correlation.

System evaluation

Prior to deploy the system in a real environment, a set of volunteers were asked to play the game, to interact with all the game features, to contribute with annotations and to provide their opinion by filling a questionnaire that will give us a first insight on how people react to the game. Although the size of the testing pool (71 participants) is much smaller than what is expected to happen in a real environment, the experiment was setup so that this small group could simulate a crowdsourcing environment and enable an analysis both on the usability of the system as well as on the quality of the annotations.

Experience setup

In order to enable analysing the results with a reduced number of actors, the experiment was setup in a way that guarantees that all the players would be guided through the same workflow and content. This included inhibiting some of the functionalities of the application like enabling the player to choose the videos he wants to contribute to, watching different videos during the crowd judgment mode, etc. The implemented evaluation process automatically guided the players through three different phases:

- On a first phase, the game competition had a duration of 3 min. We have chosen four videos to present to all the players, and defined the same starting and ending timecodes. The purpose of this mode is that the players enter the maximum number of tags that describe what they are seeing and hearing.
- In a second phase, and for 2 min, the players were asked to provide their opinion on the quality of the tags created by other players, to vote on interesting and specific moments of the video, to watch a shorter/summarised version of the video and to navigate in the tag cloud that directs them to specific time stamps of the content. During this interaction mode, we have also created some pitfalls on the displayed tags: along with the already validated tags that are shown, some incorrect tags have been added, to see how players react to them. Given that the number of functionalities available in this game mode is significant, a short tutorial was presented at the beginning of the experience, showing all the available features.

- Finally, to complete the evaluation process and unlock all the game restrictions, the player was asked to fill a questionnaire composed by 15 questions. Different aspects, important for assessing the impact of the approach, were considered: usability of the user interface, interest and motivation to play, intuitiveness of the game, players' fulfilment with the available features, and will to recommend the game to other players. The majority of the questions are linear scaled from one to five, while a few are based on multiple choice and checkbox answers and others are open answers regarding some more explicit content and players' opinion. The main objective of these last questions was to collect opinions on the main usability drawbacks and most successful functionalities as well as on new functionalities that the users would like to have included. All the questions are available on Table 3.

Questionnaire has led us to adapt some of the functionalities and improve the user interface so that some of the difficulties identified by the users were solved. The open questions gave the opportunity for users to contribute with suggestions on extra functionalities to be implemented. It was also very helpful for understanding that players were still not getting the main idea of the second phase of the game, leading some users to leave the game earlier at this stage. This information guided us on upgrading the tutorial and on providing additional functionalities and information for the players during the game: shortcut keys, sounds confirming players' actions, additional stats information, etc.

Characterisation of the participants' pool and overall analysis of the results

Tag4VD game has been released and fully available to our community of contributors for about 3 weeks. Within this period, we had 71 participants that have actively interacted with the game and have at least inserted one tag. From these participants, 47 of them have also interacted with the second phase of the experiment (the crowd judgment mode) and 28 have filled the questionnaire and unlocked all the game restrictions.

Table 3 Questionnaire questions

Type	Questions
Linear scaled	How much did you enjoy playing Tag4VD? Would you play it again? Did you feel motivated to tag the videos? Would you recommend this game to your friends? Did you learn quickly how to play? The application/website is intuitive and easy to use? How efficient/motivating is the scoring system? Did you feel motivated to unlock new badges?
Open	Which aspects did you not like? Which aspects did you like most? Do you have any improvements/suggestions regarding the scoring mechanism? Do you have any suggestions for a new badge criteria? Any other suggestions, improvements or any aspects?
Checkbox and multiple choice	I've played without reading the instructions. Which features did you like most on the second part of the game?

Phase 1

Considering all the annotations introduced, we have accomplished a total amount of 1523 tags, of which 913 (60%) consists of scored tags. These contributions allowed us to create 71 validated clusters (clusters containing more than three tags) meaning that the experiment enabled indexing 71 moments of the videos. Direct access to these timestamps and navigation in the content become then more efficient. From the point of view of timecode accuracy, the analysis of the results shows that the players tend to be very precise in time when they type some tag: 92.4% of the players were awarded 200 or 100 points, which means a half-window of 2 s from the centroid point; only 4.3% and 3.7%, were granted 50 points and 10 points, respectively, corresponding to lower time precision.

Analysis shows that 42.4% of the inserted tags have only be introduced once, while 57.6% have been introduced twice, or more. These numbers could easily be improved if misspelled words had been corrected and "test tags" (those used by players as a trial of the system, like "xyz", "123", etc.) were ignored.

The major part (47.9%) of the players added 10–50 tags. A smaller portion added 1–10 tags (45.1%) and a few players have introduced more than 50 tags (7%). Based on these numbers, we get a very reasonable average number of 21 tags per user in a quite short experience.

Semantically correlating tags allows the system to extend the default tag matching. By using the tag correlation dictionary, 47 tags were validated. The main reason for this result not to be higher is related to the video content presented during the game: it shows well known places and personalities from the town of Porto and some of the top words identify monuments, places or persons that don't exist in our database that contains essentially common words.

The results also suggest that this annotation process was successful on describing correctly objects, persons, places, etc.

Phase 2

During the next phase, players were invited to interact with all the features available on the explicit crowd judgement mode (Fig. 3): provide their opinion on the quality of existing tags through a simple "like/dislike" judgment, navigate on the tag cloud, click on the most impressive moments and watch a shorter video version.

Not all the players proceeded to this stage—the pool is smaller as only 66.2% of the initial users have reached this phase. Voting on some interesting moments in the videos was performed by 48.9% of those remaining players, while judgement on the tags that have been shown during the video presentation was completed by 59.6% of the users. The analyse of the questionnaires makes us believe that these numbers can be improved in a future and more continuous interaction with the game, as it was clear that beginners faced some difficulties on acknowledging and understanding all the available functionalities.

This experiment enabled confirming that additional information collected this way can be considered a convenient methodology to discard/approve tags created during the competition mode. As the results show, 41.4% of the presented tags have received enough crowd's opinion to be considered valid or non-valid (difference between number

of agreements and disagreements must be more or equal to a given threshold to be validated—in this experiment we setup this value to 5). Pitfall tags received a consensual crowd's opinion and have been correctly discarded by the system. With this type of information, we can then complement the scoring algorithm with crowd tag judgment, and have an additional validation instrument to the annotations created during the competition mode. However, the experiment showed that when the tags are not accurate in time, opinions tends to be divided and consequently it's harder to consider them for validation.

Figure 9 summarises some of the data collected in phase 1 and phase 2. The red line represents the median value for each group of data and the boxes represent the middle half of our data (0.5IQR). From our sample, we can see how the players interact with the different features by looking at the spread and the range of the whiskers. In this analysis, we did not consider our top level player since he showed a quite different behaviour that would distort the analysis of the main population achievements. It is clear that the range associated to the number of tags is larger than the range for the "likes". This can be justified by the fact that tag or video liking tasks are easier to perform and players are promptly engaged to this feature. The difference that can be noticed in the number of likes in tags and in videos (performed in the phase 2 of the experiment) is related to the fact that tag liking is related to a task already performed in phase 1. There is then, a clear workflow that enables players understanding this functionality. This difference in behaviour will certainly get more diffuse for loyal players in a stable and long term scenario.

Phase 3

If successful in completing the two previous phases, players were showed their statistics related to the game (score, leaderboard, unlocked badges, etc.) and were invited to answer a simple questionnaire that should not take longer than 2 min to fill.

In total, we received 28 responses. Although this number is smaller than the initial pool, it still represents 40% of the total number of volunteers and the information collected was still helpful in identifying the most enjoyable features and main drawbacks. The number of gamers that opt-out has two main explanations: on one hand the game has quite a lot of different functionalities that may require some initial training and, on

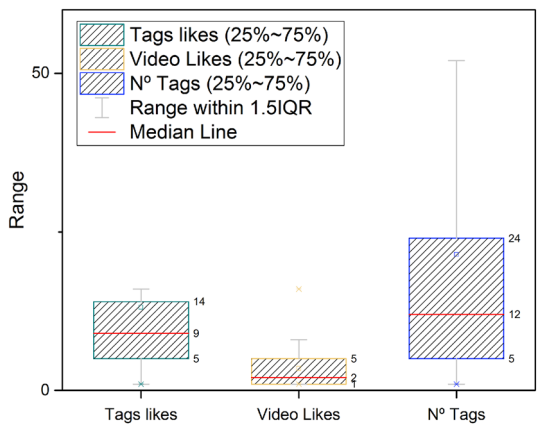

Fig. 9 User behaviour in the different game tasks

Table 4 Questionnaire results

Questions	1	2	3	4	5	Average
Did you enjoy playing Tag4VD?	2	3	9	9	5	3.43
Would you play it again?	3	3	8	11	3	3.29
Did you feel motivated to tag the videos?	0	4	8	9	7	3.68
Would you recommend this game to your friends?	3	5	9	8	3	3.11
I learned quickly how to play	2	5	5	7	9	3.57
The application is intuitive and easy to use?	2	3	5	10	8	3.68
How effective/motivating is the scoring system?	2	6	9	8	3	3.14
Did you feel motivated to unlock new badges?	4	11	7	5	1	2.57

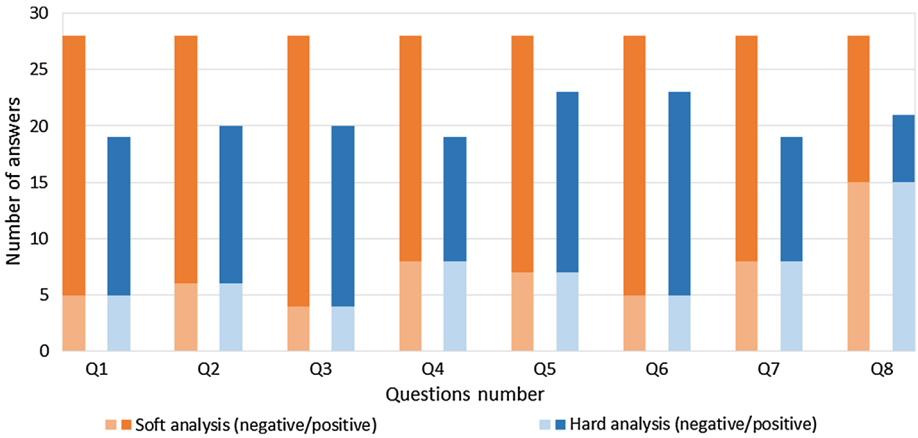

Fig. 10 Positive and negative judgments

the other hand, the experiment was setup as close to a real environment as possible—volunteers were anonymous and were not contacted nor identified.

Table 4 presents the scoring for each of the questions (number of players choosing each of the available ratings and the average value for each question) while Fig. 10 provides an additional analysis on these results, depicting the negative judgments (ratings 1 and 2) and the positive feelings (ratings 3–5). Results are very encouraging, except for last question, even if an harder analysis is done (not considering a rating of 3 as a positive score): 82% (or 74%) of the players that provided their feedback have enjoyed the game and it's features. These numbers don't of course include users that left the game before filling the form.

The motivation for unlocking new badges was a most disappointing aspect. However, this can be related to the fact that, in order to be able to guide all the users through the same game workflow, some of the functionalities were inhibited and players could not navigate freely through the game. As an example, they could neither see other players' scores nor unlocked badges until they complete this evaluation state.

One of the most consistent results shows that 47.7% of players enjoyed the feature related to quality judgement of the tags inserted by other players. Navigating on the tag cloud and clicking on the most impressive moments were the preferred features of

only 18.2% of the inquiries, while video summarization based on crucial moments of the video has 15.9% of players' appreciation.

Motivation and engagement

Productivity will depend on the motivation and enthusiasm that the gamification concepts can provide. Analysis of the answers, as well as history of participation in the testbed, enabled some conclusions on user-friendliness, usability and motivation mechanisms of the approach.

The score leaderboard, showing ranking position of a given player, demonstrated to be a good motivation factor as it was the main leveraging factor to a "battle" between our top three players. After checking up their scores, it was clear that they returned to the game and tried to improve their scores to beat opponents. Game level also demonstrated to bring enthusiasm and to contribute to enhance game performance on collecting useful information. After reaching a new level, one of the players returned back 6 times on different days. This top level player contributed with a total of 310 tags, 206 of which distinct, 118 tag judgments, unlocked 4 badges and scored 4300 points. Others players have also been motivated to come back and play more than once, demonstrating that the approach proposed had some success. In fact, half of the testbed subjects did return to the game.

Footage content itself has also proven to be a motivational factor. Some of the players mentioned that one the videos was a bit boring and the number of tags contributed to that video reflects that. This makes us believe that some information can be collected from user interaction and accomplishment of results, and correlated with classes of content or even used to find similarity between users in order to improve and optimize the set of clips to be provided for annotation. Recommendation techniques [40] can be used to improve this profile based video selection.

Conclusion

We described a web-based video annotation game which 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 scoring mechanisms take into consideration past introduced information, as well as correlated tags to enhance and improve the quality of the dataset. Different types of rewarding mechanisms that contribute for motivating good contributions were included. The metadata captured is expected to make searching and navigation in large video collections more efficient and to reduce the need for professional and expensive processes of describing content.

Analysis of the tag correlation dictionary showed that words grouped together have in fact some semantic relation. By using this approach, not only is the system able to find synonymous words, as well as other tags describing concepts that frequently occur together.

The pilot enabled a first insight on how people react to the game. Results have shown that players like to see what they have accomplished so far, which tags contributed for their scoring, and which did not. Feeling part of a community of players also demonstrated to engage users in a productive way.

As results show, our explicit user opinion feature is a good instrument to enhance the quality of the annotation by enabling the system to discard or approve wrong and correct tags. This is an innovative functionality when considering competitor systems and can be used also to increase the amount of useful metadata as validation of tags may be achieved by two distinct and complementary approaches.

Future work includes setting up a second pilot with a larger community of individuals with different backgrounds and ages with the purpose of obtaining large amounts of data. Evaluation of these annotations in comparison with professional annotations will also be done. Additional mechanisms, that include the use of content based as well as collaborative recommendation techniques, that use both explicit and implicit information to profile users, will be integrated to help on recommending the most appropriate content for each player with the objective of guaranteeing the optimization of contributions due to the increase of satisfaction.

Abbreviations

TV: television; GWAPs: games with a purpose; OCR: optical character recognition; Tag4VD: tag for video; HTML5: Hypertext Markup Language Revision 5; IQR: interquartile range.

Authors' contributions

PV was responsible for the overall conception and specification of the system. She has also been fully engaged on the setup of the experiment and on the analysis of the results. She collaborated on writing the document by defining its structure and revising the text. JPP carried out the software implementation both of the system and experiment, contributed for the analysis and evaluation of the results and first draft the manuscript. Both authors read and approved the final manuscript.

Author details

[1] INESC TEC, Campus da FEUP, Rua Dr. Roberto Frias, 378, 4200-465 Porto, Portugal. [2] ISEP/P.Porto-School of Engineering, Polytechnic of Porto, Rua Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal.

Acknowledgements

The work presented in this paper 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).

Competing interests

The authors declare that they have no competing interests.

Funding

North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF) within FourEyes, a Research Line within project "TEC4Growth – Pervasive Intelligence, Enhancers and Proofs of Concept with Industrial Impact/NORTE-01- 0145-FEDER-000020".

Received: 3 November 2016 Accepted: 20 March 2017
Published online: 12 April 2017

References

1. Marlow C, Naaman M, Boyd D, Davis M (2006) HT06, tagging paper, taxonomy, flickr, academic article, to read. In: Proceedings of the seventeenth conference on hypertext and hypermedia. ACM, pp 31–40
2. Golder SA, Huberman BA (2006) Usage patterns of collaborative tagging systems. J Inf Sci 32:198–208. doi:10.1177/0165551506062337
3. Li Q, Lu SCY (2008) Collaborative tagging applications and approaches. IEEE Multimed 15:14–21. doi:10.1109/MMUL.2008.54
4. Eggink J, Raimond Y (2013) Recent advances in affective and semantic media applications at the BBC. In: 2013 14th international workshop on image analysis for multimedia interactive services (WIAMIS), pp 1–4
5. Hildebrand M, Brinkerink M, Gligorov R et al (2013) Waisda?: Video Labeling Game. In: Proceedings of the 21st ACM international conference on multimedia. ACM, Barcelona, pp 823–826
6. Metadata Games (2015) Metadata Games - Play. Tag. Connect. http://www.metadatagames.org/. Accessed 15 Apr 2015
7. Viddler (2015) Viddler: interactive video training and practice. http://www.viddler.com/. Accessed 28 Jan 2015

8. Pinto JP, Viana P (2013) TAG4VD: a game for collaborative video annotation. In: Proceedings of the 2013 ACM international workshop on immersive media experiences. ACM, Barcelona, pp 25–28
9. Pinto JP, Viana P (2015) Using the crowd to boost video annotation processes: a game based approach. In: Proceedings of the 12th European conference on visual media production. ACM, London, pp 22:1–22:1
10. Miettinen V (2011) Digitalkoot: electrifying the finnish cultural heritage. In: Proceedings of the 4th ACM workshop on online books, complementary social media and crowdsourcing. ACM, Glasgow, pp 55–56
11. Riek LD, O'Connor MF, Robinson P (2011) Guess what? a game for affective annotation of video using crowd sourcing. In: Proceedings of the 4th international conference on affective computing and intelligent interaction-volume part I. Springer, Berlin, pp 277–285
12. OntoGames (2015) Games for semantic content creation. www.ontogame.sti2.at/games. Accessed 20 Feb 2015
13. videoTag (2015) videotag—video tagging games—a social tagging experiment. http://www.videotag.co.uk/. Accessed 25 Feb 2015
14. Brooklyn Museum—Tag! You're it! (2015) Tag! You're it!—BKM TECH. https://www.brooklynmuseum.org/community/blogosphere/2008/08/01/tag-youre-it/. Accessed 5 May 2015
15. Tiltfactor (2015) Tiltfactor - Games. http://www.tiltfactor.org/games. Accessed 5 Jun 2015
16. Siorpaes K, Hepp M (2008) Games with a purpose for the semantic web. IEEE Intell Syst 23:50–60. doi:10.1109/MIS.2008.45
17. Davis S, Burnett I, Ritz C (2009) Using social networking and collections to enable video semantics acquisition. IEEE Multimed 16(4):1. doi:10.1109/MMUL.2009.72
18. Snoek CGM, Freiburg B, Oomen J, Ordelman R (2010) Crowdsourcing rock N' roll multimedia retrieval. In: Proceedings of the 18th ACM international conference on multimedia. ACM, Firenze, pp 1535–1538
19. Mishne G (2006) AutoTag: a collaborative approach to automated tag assignment for weblog posts. In: Proceedings of the 15th international conference on world wide web. ACM, Edinburgh, pp 953–954
20. Sood S, Owsley S, Hammond K, Birnbaum L (2007) TagAssist: Automatic tag suggestion for blog posts
21. Wu B, Zhong E, Tan B, et al (2014) Crowdsourced time-sync video tagging using temporal and personalized topic modeling. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 721–730
22. Yao T, Mei T, Ngo C-W, Li S (2013) Annotation for free: video tagging by mining user search behaviour. In: Proceedings of the 21st ACM international conference on multimedia. ACM, Barcelona, pp 977–986
23. Xu P, Larson M (2014) Users tagging visual moments: timed tags in social video. In: Proceedings of the 2014 international ACM workshop on crowdsourcing for multimedia. ACM, Orlando, pp 57–62
24. Bertini M, Del Bimbo A, Torniai C (2005) Automatic video annotation using ontologies extended with visual information. In: Proceedings of the 13th annual ACM international conference on multimedia. ACM, pp 395–398
25. Moxley E, Mei T, Hua XS, et al (2008) Automatic video annotation through search and mining. In: 2008 IEEE international conference on multimedia and expo, pp 685–688
26. Larson M, Soleymani M, Serdyukov P, et al (2011) Automatic tagging and geotagging in video collections and communities. In: Proceedings of the 1st ACM international conference on multimedia retrieval. ACM, Trento, pp 51:1–51:8
27. Altadmri A, Ahmed A (2013) A framework for automatic semantic video annotation. Multimed Tools Appl 72:1167–1191. doi:10.1007/s11042-013-1363-6
28. Ferracani A, Pezzatini D, Bertini M et al (2015) A system for video recommendation using visual saliency, crowdsourced and automatic annotations. In: Proceedings of the 23rd ACM international conference on multimedia. ACM, Brisbane, pp 757–758
29. Nga DH, Yanai K (2013) Large-scale web video shot ranking based on visual features and tag co-occurrence. In: Proceedings of the 21st ACM international conference on multimedia. ACM, Barcelona, pp 525–528
30. Ballan L, Bertini M, Del Bimbo A et al (2010) Tag suggestion and localization in user-generated videos based on social knowledge. In: Proceedings of second ACM SIGMM workshop on social media. ACM, Firenze, pp 3–8
31. Chu W-T, Li C-J, Chou Y-K (2011) Tag suggestion and localization for web videos by bipartite graph matching. In: Proceedings of the 3rd ACM SIGMM international workshop on social media. ACM, pp 35–40
32. Li G, Wang M, Zheng Y-T, et al (2011) ShotTagger: tag location for internet videos. In: Proceedings of the 1st ACM international conference on multimedia retrieval. ACM, Trento, p 37:1–37:8
33. Yang Y, Yang Y, Huang Z, Shen HT (2011) transfer tagging from image to video. In: Proceedings of the 19th ACM international conference on multimedia. ACM, Scottsdale, pp 1137–1140
34. Tran H-T, Fromont E, Jacquenet F et al (2013) Unsupervised Video Tag Correction system. In: Christel Vrain André Péninou FS (ed) Extraction et gestion des connaissances (EGC'2013). Hermann-Éditions, Paris, pp 461–466
35. Baidya E, Goel S (2014) LectureKhoj: automatic tagging and semantic segmentation of online lecture videos. In: 2014 Seventh international conference on contemporary computing (IC3), pp 37–43
36. Yang H, Quehl B, Sack H (2014) A framework for improved video text detection and recognition. Multimed Tools Appl 69:217–245. doi:10.1007/s11042-012-1250-6
37. Yang H, Meinel C (2014) Content based lecture video retrieval using speech and video text information. IEEE Trans Learn Technol 7:142–154. doi:10.1109/TLT.2014.2307305
38. Chua T-S, Tang J, Hong R et al (2009) NUS-WIDE: a real-world web image database from National University of Singapore. ACM International conference on image and video retrieval
39. Yang Q, Chen X, Wang G (2008) web 2.0 dictionary. In: Proceedings of the 2008 international conference on content-based image and video retrieval. ACM, Niagara Falls, pp 591–600
40. Soares M, Viana P (2015) Tuning metadata for better movie content-based recommendation systems. Multimed Tools Appl 74:7015–7036. doi:10.1007/s11042-014-1950-1