ViTS: Video Tagging System from Massive Web Multimedia Collections

Dèlia Fernández*, David Varas*, Joan Espadaler*, Issey Masuda*, Jordi Ferreira*, Alejandro Woodward*, David Rodríguez*, Xavier Giró-i-Nieto**, Juan Carlos Riveiro* and Elisenda Bou*

*Vilynx, Palo Alto (CA) / Barcelona (Spain)
**Universitat Politècnica de Catalunya, Barcelona (Spain)

Abstract

The popularization of multimedia content on the Web has arised the need to automatically understand, index and retrieve it. In this paper we present ViTS, an automatic Video Tagging System which learns from videos, their web context and comments shared on social networks. ViTS analyses massive multimedia collections by Internet crawling, and maintains a knowledge base that updates in real time with no need of human supervision. As a result, each video is indexed with a rich set of labels and linked with other related contents. ViTS is an industrial product under exploitation with a vocabulary of over 2.5M concepts, capable of indexing more than 150k videos per month. We compare the quality and completeness of our tags with respect to the ones in the YouTube-8M dataset, and we show how ViTS enhances the semantic annotation of the videos with a larger number of labels (10.04 tags/video), with an accuracy of 80,87%. Extracted tags and video summaries are publicly available.[1]

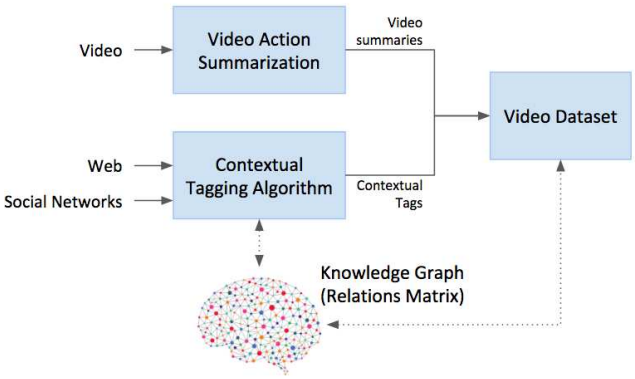


Figure 1: Scheme of ViTS. A dataset is constructed by the video summaries output from the *Video Action Summarization* block, plus the tags extracted from the *Contextual Tagging Algorithm*, which extracts tags from the information related to the video. In parallel, the *Relation Matrix* updates Knowledge Graph entities relation on real time, based on world events.

1. Introduction

During the recent years, video sharing through social media has resulted in an exponential growth of visual content available through the Internet. These video data are continuously increasing with daily recordings related to a growing number of topics and events. Manually labeling these data is extremely expensive and unfeasible in practice, therefore automatic methods for large-scale annotation are needed. Video search and indexation benefits from the use of keyword tags related to the video content, but most of the videos being shared are published without tags.

The availability of multimedia content in the Internet has also enabled the creation of large-scale video datasets, such as Sports-1M [12], YouTube-8M [1] or Kinetics [13]. Labels are available for these datasets to train and evaluate computer vision solutions in public benchmarks [9, 10, 16]. Despite the significant advances of such systems, their results are still restricted to the concepts annotated in the dataset, which typically corresponds to a single tag per video. This limitation is not acceptable for a real world application targeting a dynamic domain such as social media. For these cases, the vocabulary of labels needs an ontology-based structure, and the relations between concepts must incorporate a temporal dimension to capture the changing realities in our societies.

In this paper we present ViTS, an automatic Video Tagging System developed for large scale video tagging and summarization. The tagging algorithm is based on the ex-

[1] https://vilynx.com/research/vits_iccv2017

traction of keywords from the *contextual information*[2] associated to a video. ViTS labels are based on *Knowledge Graph* (KG) entities from Freebase [6], thus, a large and very specific vocabulary (over 2M concepts). Unlike other tagging frameworks based on closed vocabularies, our method is capable to detect real world new events, trends and concept-relation changes in real time. Having this kind of information allows for several applications such as trends detection, content-based video recommendation or indexation of large video-corpus, allowing for very specific searches of content.

Our main contribution is an online framework that crawls the web to index video documents as summaries of five seconds and a set of contextual tags. In parallel, a KG is maintained that updates over time and learns new world relations, based on the analysis of related social media. This framework is the base of an industrial product for real-time tagging of videos crawled from the Internet, which is currently indexing over 150k videos/month.

The paper is structured as follows. Section 2 provides an overview on related work for video indexing and knowledge bases. Section 3 presents a description of the KG entities, while Section 4 describes the overall system architecture. Section 5 describes how we evaluate the quality of the generated tags over a subset of YouTube-8M videos. Finally, Section 6 provides the final conclusions and points at future work.

2. Related Work

The explosion of multimedia content on the Internet generated a lot of interest on automatically annotating and indexing this content. In literature we find many Content Based Visual Retrieval (CBVR) works, which compute perceptual descriptors capable of recognizing and indexing visual content. For example, in [4] a large scale image similarity search system is presented. Other works have been studding visual semantics for large scale annotation, like [32, 35, 24]. Most recent works approach the problem with deep learning schemes which prove great performance [22, 34]. However, CBVR methods require a lot of computational resources and are sometimes not feasible for large scale and real time applications as the one targeting in this work. Moreover, large datasets are needed to train deep learning methods capable of recognizing large vocabularies of visual concepts.

In this context, a lot of effort has been applied into generating large scale datasets to train these systems: e.g. Sports-1M [12] (1M videos and ∼500 labels) for sport recognition, ActivityNet [8] (20k videos and ∼200 labels) for human activities, EventNet [37] (95k videos and 500 labels) for event-specific concepts, FCVID [11] (91k videos and 239 labels) for categories and actions, and Youtube-8M [1] (8M videos and 4.8k labels) for actions and relevant objects describing the video. Nevertheless, all these datasets but YouTube-8M, include only a few thousands of videos and the vocabulary is restricted to a few hundred of categories. Also, these vocabularies are usually very specific and not extensive to all multimedia content description and real world applications.

ViTS addresses the video indexing problem from a context-based perspective, where a light-computation solution exploit additional information associated to the video. For example, the text and metadata available in the web page where the video is embedded [14, 36, 31], or referred comments on social networks [29]. Contextual information coming from different sources requires an adaptation to unify the set of semantics used for machine tagging. In [23], a 1k concepts taxonomy called Large-Scale Concept Ontology for Multimedia (LSCOM) is presented with the purpose of standardizing multimedia annotations. Other popular models used to label multimedia content are Knowledge Bases, such as Freebase[6], WordNet[21], OpenCyc [19], Wikidata [33] or DBPedia [3]. Recent large dataset vocabulary's highly used in research are based on this kind of generic entities, e.g. ImageNet [7] and VisualGenome datasets [15] based on WordNet synsets, or YouTube-8M dataset [1] based on Google Knowledge Graph (GKG) [3] entities. This knowledge entities have many advantages compared to regular word-based vocabularies, as they standardize labels, structure the knowledge in a universal representation and model common sense. Some works are already exploring this knowledge bases to improve image classification [18] and question answering models [38]. Even the use of this knowledge bases is proving high potential, it is still a weakly explored field.

3. Knowledge Graph

The basic semantic unit generated by ViTS are *concepts* in a Knowledge Graph (KG). These concepts correspond to universal semantic representations of words. So, while *words* are dependent from a specific language, *concepts* are represented by words from different languages. The semantics associated to the concepts can be refined and updated based on the new information crawled by ViTS from the Internet. Concepts allow to merge synonymous keywords or alias under the same concept ID, i.e. the US basketball team 'Golden State Warriors' is also referred as 'San Francisco Warriors' or 'Dubs', so they all represent the same semantics. Concepts are also useful to discriminate between homonym words, i.e the word 'Queen' would map to a music band concept if appearing in a music-related context,

[2] We use 'contextual information' to refer to all the text information associated to a video URL (i.e. title, description or metadata).

[3] GKG is the extension of Freebase, since Google acquired it.

Table 1: Example of KG information saved into our database, for the tag 'New York'. Alias in bold represent the tag to show.

Wikidata ID	Freebase ID	Description	Types	Alias	Language
Q60	/m/02_286	City in New York	Place, City, Administrative Area	**New York City**	en
				The Big Apple	en
				New York	en
				NYC	en
				City of New York	en
				New Amsterdam	en
				Nueva York	es
				Ciudad de Nueva York	es
				Nova Iorque	pt
				New York	tr

while it would be mapped to 'Elizabeth II' if appearing in a context related to the British monarchy.

ViTS uses the Freebase [6] concept ID representations, which are accessible through the Google Knowledge Graph (GKG) Search API[4] and the Freebase public dumps[5]. Freebase/Wikidata dump information is integrated in our system database using Wikidata API [27]. From each concept we save in our database its *description*, *alias* (different ways how a concept can be named) in different languages, *types* extracted from GKG (e.g. 'Person', 'Place', 'City', 'Brand', 'Coorporation', 'MusicGroup'...) and its Freebase and Wikidata ID references, so we can crawl the data sources if more information is needed in a future. Sometimes concepts may not have associated *types* if they are not found in GKG API. Also, for each concept we define a tag to be displayed in every available language, which we call *tag to show*. In this work we use 'tag' to refer to the final concept translated into its tag to show. In Table 1 an example of the information saved into the database for the tag 'New York' is shown. This knowledge base results in a collection of over 2.5M KG entities, corresponding to multilingual vocabulary of over 5M words. Notice that the size of this collection constantly grows when new concepts are found on the Internet.

ViTS also tracks the relations between concepts, represented by a score value that weights the strength of the link. This score is represented in a sparse relational matrix R, of dimensions $n \times n$, where n is the total number of concepts. Each element in R represents the relation r_{ij} between two concepts c_i, c_j. The relation score r_{ij} between two concepts c_i, c_j is related to the frequency by which the two concepts co-occur in the same video:

$$r_{ij} = \frac{N_{V_{c_i} \cap V_{c_j}}}{N_{V_{c_i}}} \quad (1)$$

where $N_{V_{c_i} \cap V_{c_j}}$ is the number of videos where concept c_i has been assigned together with concept c_j, and $N_{V_{c_i}}$ is the total number of videos where concept c_i has been assigned. Notice that matrix R is not symmetric, as relations r_{ij} and r_{ji} are different.

This model allows quantifying the relations between two concepts at a low computational cost. This matrix can be updated and recalculated in real time, allowing us to quickly adapt to new events occurring in the world. Moreover, it can be time specific, taking into account only videos recorded during a temporal window. This approach is faster than word embeddings [20, 5], which have a much higher computational burden, especially when adding new concepts that would require re-training a deep learning model.

As an example of the learned relations, Figure 2 shows the 50 closest concepts to the concept 'Politics' by projecting the R matrix into a 2 dimensional space with a Multidimensional Scaling (MDS) algorithm [17]. Notice how learned relations in R generate clusters: e.g. news channels generated a cluster in the lower left corner, while politics in Spain generate a cluster in the top right corner.

4. Video Indexing

This section presents the system architecture of ViTS, which is depicted in Figure 1. The first block is the *Video Action Summarization* algorithm (4.1) that analyzes the full video using computer vision techniques to select relevant scenes. The second block is the *Contextual Tagging Algorithm* (4.2), which crawls the Internet to find keywords associated to the indexed videos and maps them to entities in the KG. The next subsections describe these two blocks in detail.

4.1. Video Action Summarization

The goal of the video action summarization block is the automatic selection of those video segments that allow a rough understanding of the semantic contents of the video.

[4] https://developers.google.com/knowledge-graph/
[5] https://developers.google.com/freebase/data

Figure 2: The 50 concepts closer to the category concept: 'Politics'. Each concept position is represented by a circle and the size of the circle represents the amount of videos tagged with this concept. Pink circles are from category concepts, while green circles are general concepts.

We consider as relevant moments those scenes that would capture viewer's attention: i.e. high action and close-ups. To this end, a set of computer vision techniques are applied to identify and extract uncorrelated clips that contain the most representative scenes, characters and objects of the video under analysis. The length of the summaries must fulfill a trade-off between being long enough to capture a whole action, but being also short enough not to mix more than one activity. In particular, ViTS builds summaries of between 3-5 seconds long.

The algorithm for video action summarization firstly segments the video into sequences with a potential of containing rich semantic information and, secondly, the relevance of each segment is estimated based on an action and object recognition engines. The details of this algorithm follows:

The algorithm for video action summarization begins with a temporal partitioning of the video into segments based on the changes of color and motion by using optical flow. The resulting segments are stabilized to compensate any camera motion. Then, objects are detected and tracked based on K-means clustering across space and time. Each object is assigned a score that takes into consideration their size, focus and position over each frame, as well as the frequency of occurrence within the whole video. In addition, the optical flow is further analyzed to compute an *action score* based on recognizing a set of predefined activities such as running or playing sports. Finally, a *Video Segment Score* is computed by summing the object and action scores, so that the N most relevant segments are kept as video summaries.

4.2. Contextual Tagging

The videos crawled by ViTS are indexed with a rich collection of tags associated to the concepts of the KG, introduced in Section 3. The tags are generated after a combination of web scraping and social media networks. The *Contextual Tagging* algorithm consists on two differentiated parts: 1) keyword extraction from Internet crawling (4.2.1) and 2) keyword mapping to KG entities (4.2.2). Figure 3 shows an scheme of the two blocks. Finally, concepts are translated to tags using its predefined *tag to show* in a given language, and are ranked for display reasons, as described in 4.2.3

4.2.1 Keyword extraction from Internet crawling

The first step towards the generation of rich tags is retrieving possible keywords available on Internet related to the indexed video. The web page where the video is embedded is parsed, so that the page title, description, text corpus and metadata are extracted from it. All Twitter posts where the video URL has been shared are also retrieved using Twitter-API [6] and parsed as well. YouTube is also crawled through YouTube Data API [7] in search of videos with the same title. If found, its related information (description and metadata) is also collected.

Once all text information is gathered, keywords are extracted using Natural Language Processing (NLP) techniques, which may differ depending on the text length and format of the extracted keywords:

[6] https://dev.twitter.com
[7] https://developers.google.com/youtube/v3/

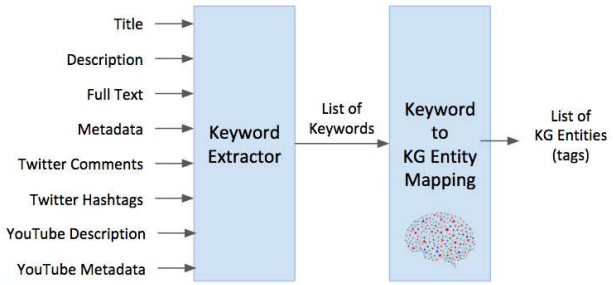


Figure 3: Scheme of the Contextual Tagging Algorithm Blocks. First blocks extracts keywords from all contextual information related to the video. Second block maps those keywords to KG entities applying context using R matrix.

- Long text (>150 words): stop words are removed and keywords are extracted using the Rapid Automatic Keyword Extraction (RAKE) algorithm [28].
- Short text (<=150 words), descriptions and titles: stop words are removed and n-grams are constructed by combining consecutive words. Each keyword is composed by n n-grams, e.g. being $n = 3$, from the title 'What reduced crime in New York City' we would get the n-grams: [*reduced, reduced crime, reduced crime New*], [*crime, crime New, crime New York*], [*New, New York, New York City*], [*York, York City*] and [*City*], where each block of n-grams is processed as an independent keyword.
- Tags from metadata: if there are already tags associated to the video, no further processing is done. These words are directly considered keywords.
- Twitter text: only the most repeated words in tweets are considered relevant and selected as keywords. The RAKE algorithm [28] is also used for this task.
- Twitter hashtags: if hashtags are composed by several words, they are split by capital letters and selected as keywords.

Finally, repeating keyword candidates are removed before generating the final list.

4.2.2 Keyword mapping to Knowledge Graph entities

The keywords extracted with the strategies presented in Section 4.2.1 must be matched with the entities of the KG introduced in Section 3.

For each keyword, we retrieve a list of *concept candidates* from the KG in our database. In particular, we search for concepts represented by similar words in the source language by using a fast text search technique which queries the keyword lexeme and returns the matching alias. If no concept candidates are found in the ViTS KG, the external Google Knowledge Graph (GKG) is used as an alternative.

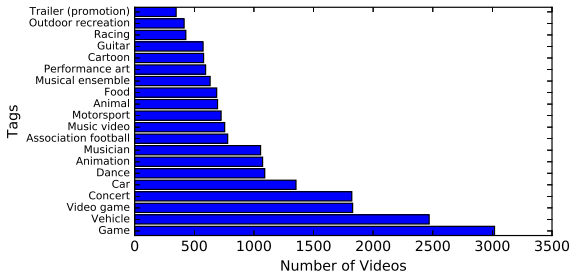
The retrieved concepts are ranked by summing two types of concept scores: an *intra-score* and an *inter-score*. Their definition is presented in the next paragraphs.

The concept *intra-score* is computed by using the information of the concept itself and it is composed of different terms. Firstly, the Levenshtein distance is computed between the keyword and the matching concept *aliases*. The Levenshtein distance corresponds to the number of deletions, insertions, or substitutions required to transform one word into another, normalized by the number of letters; i.e. if the keyword and alias are the same the distance between them is zero and it increases depending on the amount of changes needed for this two words to be the same. As we want to have a similarity score, we convert the distance into an score as $s = 1 - d$. Secondly, a concept *usability score* estimates how often the concept is used. It is computed as the linear combination of the *concept historical usability* and *concept recent usability*, being each one the ratio between the times a concept has been assigned to a video during a period of time, and all the videos processed during this same period of time. We differentiate the two scores by the time window being used: while 'historical' uses all the videos being processed by the system, 'recent' only uses a short window of time. Thirdly, a set of *Score Filters* are added to penalize or reward the score of those concepts that tend to create false positives. In our case, we work with year filters that force the matching to events in a certain year (eg. Olympic Games or Elections), as well as penalize some concepts we have manually identified as sources of false positives (eg. 'Book', 'BookSeries', 'MusicGroup', 'MusicAlbum', 'MusicComposition', 'MusicVenue', 'MovieSeries', 'Movie'). A minimum threshold is set for the concept *intra-score* which discards those concepts not reaching it.

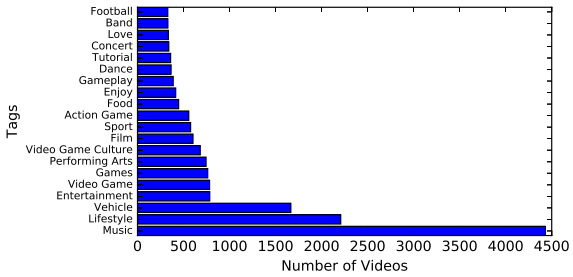
The concept *inter-score* exploits the information contained in the relational matrix R of the KG, introduced in Section 3. For each concept candidate of a given keyword (Sn_{K_i}), the relation between it and other concept candidates from other keywords is computed from matrix R by adding all relations between it and the other keyword's concept candidates, as expressed in Eq.2. Notice from the equation that relations are not computed with the concept candidates of the same keyword.

$$Cs_i = \sum_{K_j \neq K_i} R[Sn_{K_i}, Sm_{K_j}] \tag{2}$$

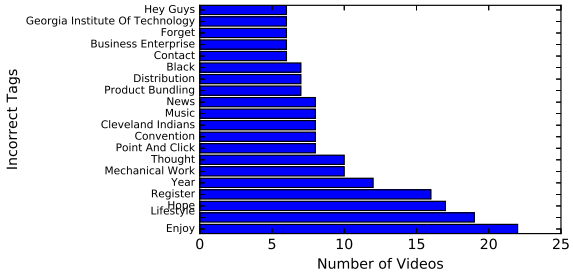
For each concept candidate, *intra-* and *inter-* scores are summed, and only those above a predefined threshold are



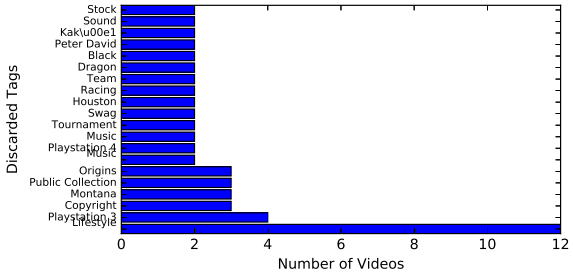
(a) The top-20 most repeated entities in the YouTube-8M subset.



(b) The top-20 most repeated entities extracted with the contextual tagging algorithm.



(c) The top-20 tags evaluated as incorrect in more videos.



(d) The top-20 tags that have been discarded in more videos.

Figure 4: Tag statistics from the AMT results.

kept.

In case of dealing with n-gram keywords, concept candidates are extracted for the n combinations of each keyword, and an score for each part of the n-gram is generated with the method explained above. Finally, the concept with the highest score is kept for each n-gram keyword.

4.2.3 Concept Ranking

Once all video concepts are extracted, they are sorted by descriptiveness and relevance for the video. This sorting is only for display purposes. We consider more relevant those tags giving specific information, i.e. name of people appearing on the video or event being shown is more relevant than general information as video categories. Following this criteria, tags are sorted using their *types*, available in the ViTS KG as explained in Section 3. Moreover, tags with equal type or with an unknown type are sorted in descendant order according to their frequency of appearance and location in the source document (title, description, social networks, etc.). Finally, concepts are translated into tags by using its *tag to show* stored in ViTS KG, as previously introduced in Section 3.

5. Experiments

The quality of the tags generated by ViTS is assessed on a subset of videos from the YouTube-8M Dataset [1]. The resulting tags from contextual information block are evaluated by human raters from the Amazon Mechanical Turk (AMT) [25] crowdsourcing platform. This Section describes the contents of the video subset used in the experiment, the statistics of the generated tags, and the assessment of their quality with AMT. The tags, summaries and video information extracted during the experiment is publicly available.

5.1. Video dataset

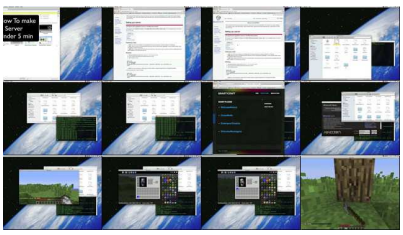
Our experiments use a subset of 13,951 videos from the public YouTube-8M video dataset [1], each of them annotated with one or more tags. Given the URL from each video, the pipeline described in Section 4.2 is applied to obtain the contextual information (title, description and metadata) that our algorithm needs to extract tags. This contextual information may include different languages, given the multilingual nature of the YouTube-8M dataset. Moreover, YouTube-8M entities are also Freebase entities, which allows a comparison between the original tags and the enhanced tags that ViTS provides.

The 13,951 videos from the subset were randomly selected and cover a large vocabulary with a wide number of topics. Figure 4a shows the distribution of videos included in the subset for the top-20 most repeated entities, translated into its *tag to show* in English. Notice how the subset has a bias towards video games, vehicles, sports and music re-

Table 2: Comparison between ViTS and YouTube-8M Tagging

ViTS	YouTube-8M
Baseball	Game
Alex Rodriguez	Arena
New York Yankees	Athlete
New York City	Baseball park
Yankee Stadium	Stadium
SportHit	Home run
Home run	

ViTS	YouTube-8M
Thomas Robinson	Basketball
Sacramento Kings	
New Jersey	
Sport	
2012 NBA Draft	



ViTS	YouTube-8M
Minecraft	Game
Video game	Minecraft
Server	
Browser extension	
Tutorial	
Download	
Video game culture	

lated entities, a distribution similar to the full YouTube-8M dataset.

5.2. Tagging Statistics

The final tags extracted by the Contextual Tagging Algorithm from the 14k videos consists on a set of 34,358 distinct KG entities. In Figure 4b we show the top-20 most repeated tags extracted by ViTS, compared to YouTube-8M's in Figure 4a. Notice a similarity on the top-level categories of the concepts: 'Music', 'Vehicles', 'Video Games', 'Food' and 'Sports'.

The average number of tags per video extracted by ViTS is 10.04, while the average number of tags in YouTube-8M dataset for the same subset of videos is 3.64. Nevertheless, in YouTube-8M tags have gone through a vocabulary construction, where all entities must have at least 200 videos in the dataset, and also only tags with visual representation are allowed, as described in [1]. In Table 2 we show a comparison of ViTS tags with respect to YouTube-8M ground truth tags for three videos. Notice the specificity of our tags and the higher quantity of tags ViTS provides.

Table 3 contains the average number of tags extracted depending on the language of the contextual information. Language is recognized by using a Wikipedia based language detection algorithm [30]. When we do not recognize the language (*null* in the table), we treat it as being English. Notice how most of the videos in the subset are in English, produces a bias on the KG Vocabulary, which is larger for English aliases. Also, relations of English topics are better learned than others. As a consequence, the average number of tags per video is higher when the contextual information is in English.

Table 3: Multilingual Tagging Statistics

Language	#Videos	Average #Tags
en	6,806	12.11
null	5,297	8.83
es	450	5.99
de	246	6.53
it	227	6.39
id	140	6.54
pt	135	4.54
nl	104	8.15
fr	90	5.68
ca	52	5.15
ro	49	6.83
tl	42	4.02
af	34	5.58
hr	30	6.06
no	28	5.92
Total	**13,951**	**10.04**

5.3. Human Rating of Generated Tags

The automatic annotations from the contextual information can be noisy and incomplete, as it is automatically generated from video title, description, metadata and user comments on social networks. The quality of the automatically generated tags was assessed by human workers from the Amazon Mechanical Turk (AMT) online platform. The tags from 1.4k randomly selected videos were shown to AMT workers, limiting the experiment to videos in English and workers located in the United States.

In each HIT (Human Intelligent Task) from AMT, three

Figure 5: Example of AMT HIT layout. On the left, video summaries are displayed in loop, together with title and video description below. On the right, the extracted tags for the video are shown for their evaluation with radio buttons.

different workers evaluated the correctness of at most 10 tags assigned to the video, ranked according to the algorithm described in Section 4.2.3. If the video had more than 10 tags associated, the additional tags were not evaluated. The video summaries, title and description from the video were shown to the worker on the user interface depicted in Figure 5. Workers were asked to decide if the tags were correct based on that information. For each tag, the worker was asked to select one of these options: *Correct, Incorrect, Do not know*. The '*Do not know*' option was added because tags may be sometimes very specific and difficult to recognize by a non-expert rater, but should not be considered incorrect for this reason. An answer was accepted when at least two workers agreed on it. If all three workers voted for the same option, we refer to it as 'absolute correct'. In case of complete disagreement, or if workers vote for majority the '*Do not know*' option, the tag is discarded. Tags extracted by ViTS that also appear in YouTube-8M ground truth were considered 'absolute correct'. Thus, these tags were not shown to the workers, but are accounted in the provided results.

Table 4 provides the accuracy results. We obtained a correctness of 77.81% of the tags evaluated, with a 77.31% of this tags with 'absolute correctness' (agreement of all 3 human raters or already in YouTube-8M annotations). Note that typical inter-rater agreement on similar annotation tasks with human raters is also around 80% [2, 26], so the accuracy of these labels is comparable to (non-expert) human-provided labels.

We also analyzed the most repeated errors and uncertain tags. Figure 4 shows the top-20 tags with more occurrences, evaluated as incorrect or discarded. Notice that many of these tags are too generic concepts, such as 'Lifestyle' or 'Music', which are often found on automatically generated metadata. Also, most of the incorrect tags are abstract concepts, like 'Enjoy', 'Hope', 'Year' or 'Thought', that are often found on contextual information but are not descriptive nor relevant to the video. Moreover, we found some

Table 4: Tag Quality Evaluation

#Videos	#Tags Total	Accuracy
1,400	14,024	80.87%

% Correct	% Incorrect	% Discarded
77.81%	18.27%	3.90%

incorrect tags caused by repeated errors on the mapping from keywords to KG entities, such as 'Georgia Institute of Technology' coming from the keyword 'technology', 'Trip Tucker' coming from 'trip' or 'Head of Mission' coming from 'cmd' or 'com'.

6. Conclusions and Future Work

This paper has introduced ViTS, an industrial Video Tagging System which generates tags based on information crawled from the Internet and learns relations between concepts. The core of the system is a knowledge base that is constantly updated to capture the dynamics of the indexed concepts.

ViTS was tested on a subset of videos from the YouTube-8M dataset. The tags generated by ViTS were highly graded by human users exposed to a visual summary of the video and its metadata. The accuracy of 80.87% is comparable to the inter-annotator agreement of (non-expert) humans in the task of semantic annotation. This high quality, combined with its capability of capturing not-only visual concepts, shows the capability of ViTS as a rich video indexing system. Moreover, experiment results on Youtube-8M are publicly available.

The presented tagging system shows how contextual data is a powerful source of information when indexing web videos. Exploiting the relations between concepts allows generating a rich set of tags with a light computation, desirable when addressing a web scale indexing. However, content-based techniques could also extend our content-based tags. Our future work will address exploiting these tags as weak labels for computer vision and audio processing deep models, which have been shown impressive recognition performances in the recent years.

Acknowledgments

Dèlia Fernàndez is funded by contract 2017-DI-011 of the Industrial Doctorate Programme of the Government of Catalonia. This work was partially supported by the Spanish Ministry of Economy and Competitivity under contracts TEC2013-43935-R and TEC2016-75976-R, and supported by grant 2014-SGR-1421 by the Government of Catalonia, and the European Regional Development Fund (ERDF).

References

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 1, 2, 6, 7

[2] R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008. 8

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. *The semantic web*, pages 722–735, 2007. 2

[4] M. Batko, F. Falchi, C. Lucchese, D. Novak, R. Perego, F. Rabitti, J. Sedmidubsky, and P. Zezula. Building a web-scale image similarity search system. *Multimedia Tools and Applications*, 47(3):599–629, 2010. 2

[5] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016. 3

[6] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM, 2008. 2, 3

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 2

[8] B. G. Fabian Caba Heilbron, Victor Escorcia and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 2

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[10] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 1

[11] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang. Exploiting feature and class relationships in video categorization with regularized deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2

[12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 1, 2

[13] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[14] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 631–640. ACM, 2007. 2

[15] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 2

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[17] J. B. Kruskal and M. Wish. *Multidimensional scaling*, volume 11. Sage, 1978. 3

[18] K. Marino, R. Salakhutdinov, and A. Gupta. The more you know: Using knowledge graphs for image classification. *arXiv preprint arXiv:1612.04844*, 2016. 2

[19] C. Matuszek, J. Cabral, M. J. Witbrock, and J. DeOliveira. An introduction to the syntax and content of cyc. In *AAAI Spring Symposium: Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, pages 44–49, 2006. 2

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 3

[21] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2

[22] M. Mühling, N. Korfhage, E. Müller, C. Otto, M. Springstein, T. Langelage, U. Veith, R. Ewerth, and B. Freisleben. Deep learning for content-based video retrieval in film and television production. *Multimedia Tools and Applications*, pages 1–26. 2

[23] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE multimedia*, 13(3):86–91, 2006. 2

[24] G. d. Oliveira Barra. Livre: A video extension to the lire content-based image retrieval system. 2015. 2

[25] G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running experiments on amazon mechanical turk. 2010. 6

[26] R. Passonneau, N. Habash, and O. Rambow. Inter-annotator agreement on a multilingual semantic annotation task. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1951–1956, 2006. 8

[27] T. Pellissier Tanon, D. Vrandečić, S. Schaffert, T. Steiner, and L. Pintscher. From freebase to wikidata: The great migration. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1419–1428. International World Wide Web Conferences Steering Committee, 2016. 3

[28] S. Rose, D. Engel, N. Cramer, and W. Cowley. Automatic keyword extraction from individual documents. *Text Mining*, pages 1–20, 2010. 5

[29] B. Shevade, H. Sundaram, and L. Xie. Modeling personal and social network context for event annotation in images. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 127–134. ACM, 2007. 2

[30] N. Shuyo. Language detection library for java, 2010. 7

[31] E. Spyrou and P. Mylonas. Analyzing flickr metadata to extract location-based information and semantically organize its photo content. *Neurocomputing*, 172:114–133, 2016. 2

[32] D. Tsai, Y. Jing, Y. Liu, H. A. Rowley, S. Ioffe, and J. M. Rehg. Large-scale image annotation using visual synset. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 611–618. IEEE, 2011. 2

[33] D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014. 2

[34] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li. Deep learning for content-based image retrieval: A comprehensive study. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 157–166. ACM, 2014. 2

[35] X.-J. Wang, L. Zhang, M. Liu, Y. Li, and W.-Y. Ma. Arista-image search to annotation on billions of web photos. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2987–2994. IEEE, 2010. 2

[36] X. Wu, C.-W. Ngo, A. G. Hauptmann, and H.-K. Tan. Real-time near-duplicate elimination for web video search with content and context. *IEEE Transactions on Multimedia*, 11(2):196–207, 2009. 2

[37] G. Ye, Y. Li, H. Xu, D. Liu, and S.-F. Chang. Eventnet: A large scale structured concept library for complex event detection in video. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 471–480. ACM, 2015. 2

[38] Y. Zhu, A. Fathi, and L. Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *European conference on computer vision*, pages 408–424. Springer, 2014. 2