Proc. 25th Int. Conf. on Multimedia Modeling (MMM 2019), Springer LNCS vol. 11295. Author's accepted version.
The final publication is available at Springer via https://doi.org/10.1007/978-3-030-05710-7_12

Multimodal Video Annotation for Retrieval and Discovery of Newsworthy Video in a News Verification Scenario

Lyndon Nixon[1], Evlampios Apostolidis[2,3], Foteini Markatopoulou[2], Ioannis Patras[3], and Vasileios Mezaris[2]

[1] MODUL Technology GmbH, Vienna, Austria
nixon@modultech.eu
[2] Centre for Research and Technology Hellas, Thermi-Thessaloniki, Greece
{apostolid, markatopoulou, bmezaris}@iti.gr
[3] School of EECS, Queen Mary University of London, London, UK
i.patras@qmul.ac.uk

Abstract. This paper describes the combination of advanced technologies for social-media-based story detection, story-based video retrieval and concept-based video (fragment) labeling under a novel approach for multimodal video annotation. This approach involves textual metadata, structural information and visual concepts - and a multimodal analytics dashboard that enables journalists to discover videos of news events, posted to social networks, in order to verify the details of the events shown. It outlines the characteristics of each individual method and describes how these techniques are blended to facilitate the content-based retrieval, discovery and summarization of (parts of) news videos. A set of case-driven experiments conducted with the help of journalists, indicate that the proposed multimodal video annotation mechanism - combined with a professional analytics dashboard which presents the collected and generated metadata about the news stories and their visual summaries - can support journalists in their content discovery and verification work.

Keywords: News video verification, Story detection, Video retrieval, Video fragmentation, Video annotation, Video summarization.

1 Introduction

Journalists and investigators alike are increasingly turning to online social media to find media recordings of events. Newsrooms in TV stations and online news platforms make use of video to illustrate and report on news events, and since professional journalists are not always at the scene of a breaking or evolving story, it is the content posted by users that comes into question. However the rise of social media as a news source has also seen a rise in fake news - the spread of deliberate misinformation or disinformation on these platforms. Images and videos have not been immune to this, with easy access to software to tamper with and modify media content leading to deliberate fakes, although fake media

Proc. 25th Int. Conf. on Multimedia Modeling (MMM 2019), Springer LNCS vol. 11295. Author's accepted version.
The final publication is available at Springer via https://doi.org/10.1007/978-3-030-05710-7_12

can also be the re-posting of a video of an earlier event, with the claim that it shows a contemporary event.

Our InVID project [4] has the goal to facilitate journalists in identifying online video posted to social networks claiming to show news events, and verifying that video before using it in reporting (Section 2). This paper presents our work on newsworthy video content collection and multimodal video annotation that allows fine-grained (i.e. at the video-fragment-level) content-based discovery and summarization of videos for news verification, since the first step for any news verification task must be finding the relevant (parts of) online posted media for a given news story. We present the advanced techniques developed to detect news stories in a social media stream (Section 3), retrieve online posted media from social networks for those news stories (Section 4), fragment each collected video into visually coherent parts and annotate each video fragment based on its visual content (Section 5). Following, we explain how these methods are pieced together forming a novel multimodal video annotation methodology, and describe how the tailored combination of these technologies in the search and browsing interface of a professional dashboard can support the fine-grained content-driven discovery and summarization of the collected newsworthy media assets (Section 6). Initial experiments with journalists indicate the added value of the generated visual summaries (Section 7) for the in-time discovery of the most suitable video fragments to verify and present a story. Section 8 concludes the work reported in this paper.

2 Motivation

Even reputable news agencies have been caught out, posting images or videos in their reporting of news stories that turn out later to have been faked or falsely associated. It is surely not the last time fake media will end up being used in news reporting, despite the growing concerns about deliberate misinformation being generated to influence social and political discussion. Journalists are under time-pressure to deliver, and often the only media illustrating a news event is user-provided and circulating on social networks. The current process for media verification is manual and time-consuming, pursued by journalists who lack the technical expertise to deeply analyze the online media. The in-time identification of media posted online, which (claim to) illustrate a (breaking) news event is for many journalists the foremost challenge in order to meet deadlines to publish a news story online or fill a news broadcast with content.

Our work is part of an effort to provide journalists with a semi-automatic media verification toolkit. While the key objective is to facilitate the quicker and more accurate determination of whether an online media file is authentic (i.e. it shows what it claims to show without modification to mislead the viewer), a necessary precondition for this is that the user can identify the appropriate candidates for verification from the huge bulk of media content being posted online continually, on platforms such as Twitter, Facebook, YouTube or DailyMotion.

[4] https://www.invid-project.eu/

Proc. 25th Int. Conf. on Multimedia Modeling (MMM 2019), Springer LNCS vol. 11295. Author's accepted version.
The final publication is available at Springer via https://doi.org/10.1007/978-3-030-05710-7_12

The time taken by journalists to select a news story and find online media for that news story, directly affects the time they have remaining to additionally conduct the verification process on that media. Hence, the timely identification of news stories, as well as the accurate retrieval of candidate media for those stories, is the objective of this work. Given that the journalistic verification and re-use (in reporting) of content only needs that part of the content which shows the desired aspect of the story (possibly combining different videos to illustrate different aspects), the fragmentation of media into conceptually distinct and self-contained fragments is also relevant.

The state of the art in this area would be e.g. using TweetDeck to follow a set of news-relevant terms continually (e.g. "just happened", "explosion" etc.), while manually adding other terms and hashtags when a news story appears (e.g. by tracking news tickers or Twitter's trending tags). Once media is found it has to be examined (e.g. a video played through) to check its content and identify the part(s) of interest. This is all still very time-consuming and prone to error (e.g. missing relevant media). Hence, this paper reports on our contribution - in the context of the journalistic workflow - to the content-based discovery and summarization of video shared on social platforms. The proposed system (see Fig. 1) combines advanced techniques for news story detection, story-driven video retrieval and video fragmentation and annotation, into a multimodal video annotation process which produces a rich set of annotations that facilitate the fine-grained discovery and summarization of video content related to a news story, through the interactive user interface of a professional multimodal analytics dashboard. This system represents an important step beyond the state of the art in the journalistic domain, as validated by a user evaluation.

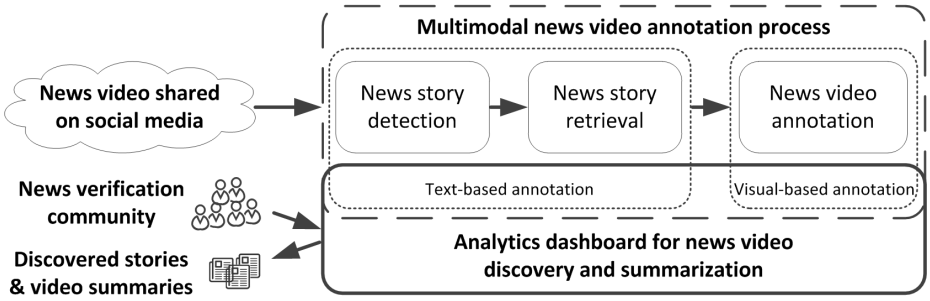

Fig. 1. The overall architecture of the proposed system.

The following Sections 3 to 5 describe the different components of the system and report on their performance using appropriate datasets. Section 6 explains how these components are pieced together into the proposed multimodal video annotation mechanism and the multimodal analytics dashboard to enable the content-based retrieval, discovery and summarization of the collected data. Finally, Section 7 reports the case-driven evaluations about the efficiency of the proposed system, made by journalists from two news organizations.

Proc. 25th Int. Conf. on Multimedia Modeling (MMM 2019), Springer LNCS vol. 11295. Author's accepted version.
The final publication is available at Springer via https://doi.org/10.1007/978-3-030-05710-7_12

3 News Story Detection

The developed story detection algorithm uses content extracted from the Twitter Streaming API, modelling each tweet as a bag of keywords, clustering keywords in a weighted directed graph and using cluster properties to merge/split the clusters into distinct news stories [5]. The output is a ranked list of news stories labelled by the three most weighted keywords in the story cluster and explained by (up to) ten most relevant documents (i.e. videos) in our collection related to that story (based on keyword matching). Fig. 2 shows the story detection in the dashboard: a story is given a label and the most relevant documents are shown underneath.

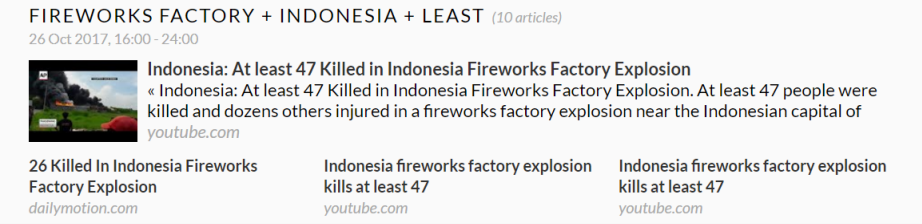

Fig. 2. Presentation of a story in the dashboard.

Guided by the fact that a human evaluator is required to assess the quality and correctness of the stories, we manually evaluated the performance of the story detection algorithm considering two factors: quality and correctness. For this, a daily assessment of the top-10 stories detected from our Twitter Accounts stream used for news story detection, was conducted by the lead author on May 28 to May 30, 2018. For each story, we evaluated both the defined label (does it meaningfully refer to a news story? does it reflect a single story or multiple stories?) and the documents presented for that story (do they relate to the story represented by its label?). From this evaluation, we could combine our insights into four metrics which we can compare across sources and days:

- Correctness: indicates if the generated clusters correctly relate to newsworthy stories;
- Distinctiveness: evaluates how precisely each individual cluster relates to an individual story;
- Homogeneity: examines if the documents in the cluster are only relevant to the newsworthy stories represented by the cluster;
- Completeness: assesses the relevance of the documents in the cluster with a single, distinct news story.

The results reported in Table 1 demonstrate that our method performs almost perfectly on providing newsworthy events and separating them distinctly. Sample size was n=10 for the story metrics (correctness and distinctiveness) and n=100 for the story document metrics (homogeneity and completeness).

Proc. 25th Int. Conf. on Multimedia Modeling (MMM 2019), Springer LNCS vol. 11295. Author's accepted version.
The final publication is available at Springer via https://doi.org/10.1007/978-3-030-05710-7_12

Table 1. Story detection comparison for May 2018.

	Correctness	Distinctiveness	Homogeneity	Completeness
May 28, 2018	1	1	0.94	0.9
May 29, 2018	1	1	0.95	0.95
May 30, 2018	1	0.8	0.98	0.98
Three day average	**1**	**0.93**	**0.96**	**0.94**

4 News Video Retrieval

Given a set of detected news stories, the next step is to select from social networks candidate videos which claim to illustrate those stories. Multimedia Information Retrieval is generally evaluated against a finite video collection that is extracted from the social network data using some sampling technique, and then feature-based or semantic retrieval is tested on this finite collection. This is different from our goal, which is to maximize the relevance of the media documents returned for our Web-based queries, based on the news story detection.

On one hand, classical recall cannot be used since the indication of the total number of relevant documents on any social media platform at any time for one query is not possible. On the other hand, we should consider whether success in information retrieval only occurs if and only if the retrieved video is relevant to the story represented by the query, or if any newsworthy video being retrieved can be considered a sign of success. Indeed, videos related to a news story will keep being posted for a longer time after the news story initially occurred, and those videos can still be relevant for discovery in a news verification context. Thus, queries which reference keywords that persist in the news discussion, (e.g. "Donald Trump"), are likely to return other videos which are not relevant to the current story, but still reference an earlier newsworthy story. Classical "Precision" can indicate how many of the retrieved videos are relevant to the query itself, yet low precision may hide the fact that we still collect a high proportion of newsworthy video content. Still, it acts as an evaluation of the quality of our query to collect media for the specific story. We choose a second metric, titled "Newsworthiness", which measures the proportion of all newsworthy video returned for a query. Since this metric includes video not directly related to the story being queried for, it evaluates the appropriateness of our query for collecting newsworthy media in general. Finally, we define a "Specificity" measure as the proportion of newsworthy video retrieved that is relevant to the story being queried for, ergo our Specificity is the Precision divided by the Newsworthiness and assesses the specificity of our query for the news story.

After experimenting with different query-construction approaches based on how the stories are represented in our system (clusters of keywords) [5], we propose the story labels (top-3 weighted keywords in the cluster) as the basis for querying social networks for relevant videos. We used the story labels from the top-10 stories from Twitter Accounts in the dashboard for the period May 28-30, 2018 as conjunctive query inputs. We tested the results' relevance by querying

Proc. 25th Int. Conf. on Multimedia Modeling (MMM 2019), Springer LNCS vol. 11295. Author's accepted version.
The final publication is available at Springer via https://doi.org/10.1007/978-3-030-05710-7_12

the YouTube API, using the default result sort by relevance, and measuring "Precision at N", where N provides the cut-off point for the set of documents to evaluate for relevance. In line with the first page of search results, a standard choice in Web Search evaluation, we chose $n = 20$ for each story. Since each day we use the top-10 stories, the retrieval is tested on a sample of 200 videos. In Table 2 we compare the results from last year on 13 June 2017 (which acts as a benchmark for our work [5]) and the results for the aforementioned dates in May 2018 (and their average). It can be seen that our Specificity value has increased considerably, meaning that when we make a query for a newsworthy story we are more likely to only get videos that are precisely relevant to that story, than video of any newsworthy story. So, while Newsworthiness has remained more or less the same (the proportion of newsworthy video being collected into the developed platform is probably still around 80% for YouTube), our Precision at N value - that the collected video is precisely about the news story we detected - shows an over 20% improvement. Our news video retrieval technique collects documents from a broad range of stories with a consistent Precision of around 0.76 and Specificity of around 0.94, which indicate that document collection is well-balanced across all identified news stories. Since this video retrieval mechanism has been integrated in the InVID dashboard we add around 4000 - 5000 new videos per day, based on queries for up to 120 detected unique news stories.

Table 2. Our social media retrieval tested on the new story labels.

	13 June 2017	28 May 2018	29 May 2018	30 May 2018	2018 avg.
Precision	**0.54**	0.79	0.7	0.79	**0.76**
Newsworthiness	**0.82**	0.85	0.74	0.84	**0.81**
Specificity	**0.64**	0.93	0.95	0.94	**0.94**
F-score	**0.59**	0.82	0.72	0.81	**0.78**

5 News Video Annotation

Every collected video is analysized by the video annotation component which produces a set of human-readable metadata about the videos' visual content at the fragment-level. This component segments the video into temporally and visually coherent fragments and extracts one representative keyframe for each fragment. Following, it annotates each segment with a set of high-level visual concepts after assessing their occurrence in the visual content of the corresponding keyframe. The produced fragment-level conceptual annotation of the video can be used for fine-grained concept-based video retrieval and summarization.

The temporal segmentation of a video into its structural parts, called shots, is performed using a variation of [1]. The visual content of the frames is represented with the help of local (ORB [9]) and global (HSV histograms) descriptors. Then, shot boundaries are detected by assessing the visual similarity between consecutive and neighboring video frames and comparing it against experimentally defined thresholds and models that indicate the existence of abrupt and

Proc. 25th Int. Conf. on Multimedia Modeling (MMM 2019), Springer LNCS vol. 11295. Author's accepted version.
The final publication is available at Springer via https://doi.org/10.1007/978-3-030-05710-7_12

gradual shot transitions. These findings are re-evaluated using a pair of dissolve and wipe detectors (based on [12] and [11] respectively) that filter-out wrongly detected gradual transitions due to swift camera and/or object movement. The final set of shots is formed by the union of the detected abrupt and gradual transitions, and each shot is represented by its middle frame. Evaluations using the experimental setup of [1], highlight the efficiency of this method. Precision and Recall are equal to 0.943 and 0.941 respectively, while the needed processing time (13.5% of video duration, on average) makes the analysis over 7 times faster than real-time processing. These outcomes indicate the ability of this method to process large video collections in a highly-accurate and time-efficient manner.

When dealing with raw, user-generated videos the shot-level fragmentation is too coarse and fails to reveal information about their structure. A decomposition into smaller parts, called sub-shots, is needed to enable fine-grained annotation and summarization of their content. Guided by this observation, we define a sub-shot as an uninterrupted sequence of frames having only a small and contiguous variation in their visual content. The algorithm (denoted as "DCT") represents the visual content of frames using a 2D Discrete Cosine Transform and assesses their visual resemblance using cosine similarity. The computed similarity scores undergo a filtering process to reduce the effect of sudden, short-term changes in the visual content; the turning points of the filtered series of scores signify a change in the similarity tendency and therefore a sub-shot boundary. Finally, each defined sub-shot is represented by the frame with the most pronounced change in the visual content. The performance of this method was evaluated using a relevant dataset [5] and compared against other approaches, namely: a) a method similar to [7], which assesses the visual similarity of video frames using HSV histograms and the x^2 metric (denoted as "HSV"); b) an approach from [2], which estimates the frame affinity and the camera motion by extracting and matching SURF descriptors (denoted as "SURF"); and c) an implementation of the best performing technique of [2], that estimates the frame affinity and the camera motion by computing the optical flow (denoted as "AOF"). The evaluation outcomes (see Table 3) indicate the DCT-based algorithm as the best trade-off between accurate and fast analysis. The analysis is over 30 times faster than real-time processing and results in a rich set of video fragments that can be used for fine-grained video annotation.

Table 3. Performance evaluation of the used sub-shot segmentation algorithm.

	DCT method	HSV method	SURF method	AOF method
Precision	0.22	**0.44**	0.36	0.27
Recall	**0.84**	0.11	0.29	0.78
F-Score	0.36	0.18	0.33	**0.40**
Proc. time (x video length)	**0.03**	0.04	0.56	0.08

[5] Publicly available at https://mklab.iti.gr/results/annotated-dataset-for-sub-shot-segmentation-evaluation/

Proc. 25th Int. Conf. on Multimedia Modeling (MMM 2019), Springer LNCS vol. 11295. Author's accepted version.
The final publication is available at Springer via https://doi.org/10.1007/978-3-030-05710-7_12

The concept-based annotation of the defined video fragments is performed using a combination of deep learning methods (presented in [8] and [4]), which evaluate the appearance of 150 high-level concepts from the TRECVID SIN task [6] in the visual content of the corresponding keyframes. Two pre-trained ImageNet [10] deep convolutional neural networks (DCNNs), have been fine-tuned (FT) using the extension strategy of [8]. Similar to [4], the networks' loss function has been extended with an additional concept correlation cost term; giving a higher penalty to pairs of concepts that present positive correlations but have been assigned with different scores, and the same penalty to pairs of concepts that present negative correlation but have not been assigned with opposite scores. The exact instantiation of the used approach is as follows: Resnet1k-50 [3] extended with one extension FC layer with size equal to 4096 and GoogLeNet [13] trained on 5055 ImageNet concepts [10], extended with one extension FC layer of size equal to 1024. During semantic analysis each selected keyframe is forward propagated by each of the FT networks described above; each network returns a set of scores that represent its confidence regarding the concepts' occurrence in the visual content of the keyframe. The scores from the two FT networks for the same concept are combined in terms of arithmetic mean. The performance of this technique has been evaluated in terms of MXinfAP (Mean Extended Inferred Average Precision) on the TRECVID SIN 2013 dataset. The employed method achieved a MXinfAP score equal to 33.89%, thus proven to be a competitive concept-based annotation approach. After analyzing the entire set of extracted keyframes the applied method produces a fragment-level conceptual annotation of the video which enables concept-based video retrieval and summarization.

6 Concept-based Summaries

The functionality of the aforementioned technologies is combined with a professional multimodal analytics dashboard, in a way tailored to the needs of journalists who want to quickly discover the most suitable newsworthy video content shared on social media platforms. The news story detection and retrieval components of the dashboard enable the creation of story-related collections of newsworthy videos. Then, a multimodal video annotation approach takes place for every collected video. It produces a set of text-based annotations at the video-level according to the associated metadata, and a set of concept-based annotations that represent the visual content of the video at the fragment-level. The dashboard provides a user interface to the aforementioned collected and generated metadata for every newsworthy video that is inserted to the system, allowing the users to quickly find (parts of) online video relevant to a news story. Based on the outcomes of the applied multimodal video annotation process the collected video content can be browsed at different levels of granularity - i.e. the story-level (groups of videos related to a story), the document-level (a single video about a story) and the fragment-level (a particular video fragment showing an event of interest) - through textual queries or by matching visual attributes with the visual content of the documents. In the latter case the user is able to

Proc. 25th Int. Conf. on Multimedia Modeling (MMM 2019), Springer LNCS vol. 11295. Author's accepted version.
The final publication is available at Springer via https://doi.org/10.1007/978-3-030-05710-7_12

retrieve and discover a summary of a video document that includes only the fragments that relate to a selected set of (event-related) visual concepts.

As a result, we have the possibility to offer concept-based summaries of videos of a news story. For example, a journalist looking for a video of the fire at Trump Tower, New York City (Jan 8, 2018) can text search for "trump tower fire" and find videos in that time period which reference this in their textual metadata (title + description). However, the videos do not necessarily show the fire (one observed phenonema on YouTube has been putting breaking news story titles into video descriptions as "clickbait" for views) and those who do, may be longer and the journalist still needs to search inside the video for the moment the fire is shown. With the visual concept search, a text search on "trump tower" can be combined with a concept-based search on EXPLOSION_FIRE. This would return a list of videos with their fragments where the fire is visible. The retrieved video fragments of each video, which form a concept-related visual summary of the video, are represented by their keyframes in the user interface of the dashboard, but can be also played back allowing the user to watch the entire fragment. Below in Fig. 3 we show a conceptual summary for a video returned for the story of the Thai cave rescue story with the visual concept of "Swimming".

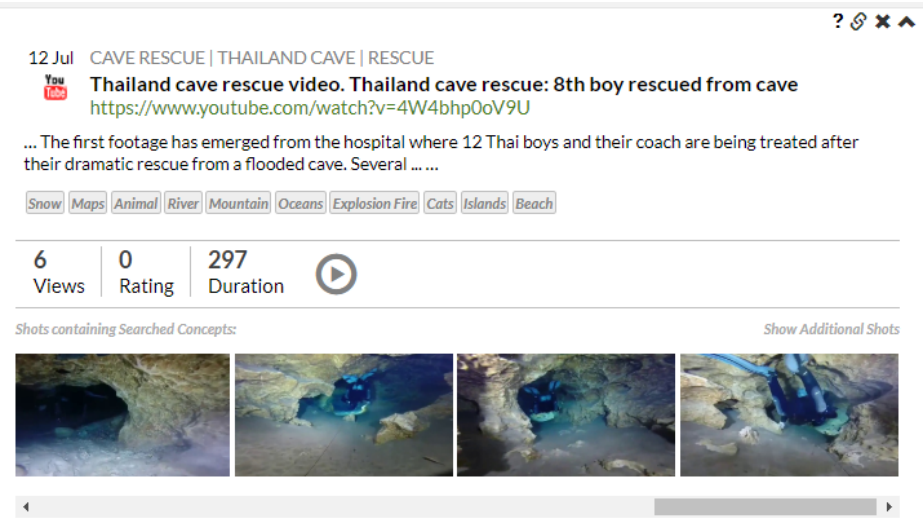

Fig. 3. Video fragments with the "Swimming" concept for a "Thai cave rescue" video.

7 Evaluation

A user evaluation is necessary to assess the efficiency of our story detection, video retrieval and summarization technologies. As opposed to the evaluation of the accuracy of these techniques through comparison with some benchmark, the

Proc. 25th Int. Conf. on Multimedia Modeling (MMM 2019), Springer LNCS vol. 11295. Author's accepted version.
The final publication is available at Springer via https://doi.org/10.1007/978-3-030-05710-7_12

"best" solution for a journalist is less the automatic achievement of 100% accuracy, and more the successful discovery in less time of relevant video material for verification. In our case, we have a clear use case for this and involve journalists in determining if the techniques, as provided through the dashboard, meet their expectations. We asked two journalists with verification expertise, one in Agence France Presse (AFP) and the other in Deutsche Welle (DW), to perform a number of queries for video documents associated to a news story detected by our story detection algorithm. For each of the selected stories, they used the multimodal analytics dashboard to select the top video documents associated with that story and to find a video snippet showing the desired visual component of the story (i.e. something which could be used in a news report after verification). In both cases, they can explore video documents either by conducting a textual search over the document set or generating a concept-based summary using the most appropriate visual concept. Finally, we surveyed which of the results were more useful for their journalistic work. Whereas we involved just two journalists in this initial evaluation, the evaluation methodology is based on previous discussions with the organizations and thus reflects adequately a real-life work scenario for journalists in general.

The journalist from AFP chose the story of the Thailand cave rescue, a story which was detected in the dashboard with the label CAVE + RESCUE + THAILAND. The system had retrieved 1317 videos for this story in the period 9-12 July. The visual component of interest in this story was a snippet of the divers in the cave undertaking the rescue operation. Firstly, we chose the term "divers" for further searching the 1317 videos, and got 142 matching videos which are ordered in the dashboard by a relevance metric. The journalist viewed seven videos until selecting an appropriate video snippet. In total, five video fragments were selected for viewing during the browsing process (from four distinct videos; the other three videos were previously discarded as irrelevant). The time needed was 8 minutes, mainly as the videos were presented in their entirety (all fragments) and time was spent on looking at videos which did not have any relevant content. Then, we chose as a visual concept that of "swimming", since in the context of the story only the divers would be swimming in the Thai cave. This returned 79 results. The journalist viewed seven videos until selecting an appropriate video snippet. In total, only two video fragments were viewed while browsing as they were now already filtered by the selected concept, and the displayed keyframes were already indicative of whether the video fragment actually showed divers in the cave or something else. The time needed was now 4 minutes as the videos were more relevant and could be more quickly browsed, as only potentially matching fragments were being shown in the dashboard. As a side note, at least one video could be considered as "fake" - it differed significantly from the other observed footage, contained other diving footage and was entitled with the Thai cave rescue story as "clickbait" - something that other verification tools developed in our project (e.g. the InVID Plug-in [14]) could help establish.

The journalist from DW tested on 21 July 2018 when one of the main news stories in the dashboard was the tragic sinking of a "duck boat" in Missouri,

Proc. 25th Int. Conf. on Multimedia Modeling (MMM 2019), Springer LNCS vol. 11295. Author's accepted version.
The final publication is available at Springer via https://doi.org/10.1007/978-3-030-05710-7_12

USA. The story was detected with the label BOAT + DUCK + MISSOURI. The system had retrieved 292 videos for this story in the period 21-22 July. The visual component of interest in this story was a snippet of the boat sinking in the lake. Firstly, we chose the term "sinking" to search the collection of 292 videos, getting 24 matching videos. In the results ordered by relevance, we had to discount three videos which were related with "sinking" but not with the Missouri duck boat story. The journalist viewed ten videos until an appropriate video snippet was selected. In total, five video fragments were viewed from three distinct videos. The time needed was 6 minutes, as it took some time to get to a video with an appropriate fragment. Then, we chose the visual concept of "boat-ship", since this would be the visual object doing the "sinking" in the video. This returned 137 videos. Actually the very first video in this case provided two usable fragments showing the boat, and thus less than one minute was actually needed in this case. However, we continued to analyse the remaining results (10 most relevant). A set of eleven fragments representing six videos was viewed as relevant, and two further usable fragments were identified (all coming from the same source, which is presumably a claimed user-recording of the sinking boat).

The evaluations, focused on the journalistic workflow, established that while both the text-based and concept-based searches could adequately filter video material for a detected news story, when searching for a specific visual event in the video the concept-based search was comparatively quicker. The returned videos were more relevant content-wise, as often in textual metadata the event is described but not necessarily shown in the video. The dashboard display of matching fragments (via their keyframes) could help the journalists quickly establish if the content of the video was what they were searching for, and the filtering of the fragments in a concept-based search brings the journalists quicker to the content they would like to use. The presence of a probable "fake" among the Thai cave rescue videos acted as a reminder of why journalists want to find content quickly: the time is needed to verify the authenticity of an online video before it should be finally used in any news reporting.

8 Conclusion

This paper described a novel multimodal approach for newsworthy content collection, discovery and summarization, that facilitates journalists to quickly find online media associated with current news stories and determine which (part of) media is the most relevant to verify before potentially using it in a news report. We presented a solution to news story detection from social media streams, we outlined how the news story descriptions are used to dynamically collect associated videos from social networks, and we explained how we fragment and conceptually annotate the collected videos. Then, we discussed a combination of these methods into a novel multimodal video annotation mechanism that annotates the collected video material at different granularities. Using a professional multimodal analytics dashboard that integrates the proposed news video content collection and annotation process, we provided a proof of concept-based sum-

marization of those videos at the fragment-level. The evaluation of the proposed approach with journalists showed that the concept-based search and summarization of news videos allows journalists to find quicker the most suitable parts of the video content, for verifying the story and preparing a news report. The dashboard is now being tested with a larger sample of external journalistic users, which will provide more comprehensive insights into the value and the limitations of the work presented in this paper.

9 Acknowledgments

This work was supported by the EUs Horizon 2020 research and innovation programme under grant agreement H2020-687786 InVID.

References

1. Apostolidis, E., Mezaris, V.: Fast shot segmentation combining global and local visual descriptors. In: 2014 IEEE Int. Conf. on Acoustics, Speech and Signal Processing. pp. 6583–6587 (2014)
2. Cooray, S.H., O'Connor, N.E.: Identifying an efficient and robust sub-shot segmentation method for home movie summarisation. In: 10th Int. Conf. on Intelligent Systems Design and Applications. pp. 1287–1292 (2010)
3. He, K., Zhang, X., et al.: Deep residual learning for image recognition. In: 2016 IEEE Conf. on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
4. Markatopoulou, F., Mezaris, V., et al.: Implicit and explicit concept relations in deep neural networks for multi-label video/image annotation. IEEE Trans. on Circuits and Systems for Video Technology pp. 1–1 (2018)
5. Nixon, L.J.B., Zhu, S., et al.: Video retrieval for multimedia verification of breaking news on social networks. In: 1st Int. Workshop on Multimedia Verification (MuVer '17) at ACM Multimedia Conf. pp. 13–21. MuVer '17, ACM (2017)
6. Over, P.D., Fiscus, J.G., et al.: TRECVID 2013-An overview of the goals, tasks, data, evaluation mechanisms and metrics. In: TRECVID 2013. NIST, USA (2013)
7. Pan, C.M., Chuang, Y.Y., et al.: NTU TRECVID-2007 fast rushes summarization system. In: TRECVID Workshop on Video Summarization. pp. 74–78. ACM (2007)
8. Pittaras, N., Markatopoulou, F., et al.: Comparison of fine-tuning and extension strategies for deep convolutional neural networks. In: 23rd Int. Conf. on MultiMedia Modeling. pp. 102–114 (2017)
9. Rublee, E., Rabaud, V., et al.: ORB: An efficient alternative to SIFT or SURF. In: 2011 Int. Conf. on Computer Vision. pp. 2564–2571 (2011)
10. Russakovsky, O., Deng, J., et al.: ImageNet Large Scale Visual Recognition Challenge. Int. Journal of Computer Vision **115**(3), 211–252 (2015)
11. Seo, K., Park, S.J., et al.: Wipe scene-change detector based on visual rhythm spectrum. IEEE Trans. on Consumer Electronics **55**(2), 831–838 (2009)
12. Su, C.W., Tyan, H.R., et al.: A motion-tolerant dissolve detection algorithm. In: IEEE Int. Conf. on Multimedia and Expo. vol. 2, pp. 225–228 (2002)
13. Szegedy, C., Liu, W., et al.: Going deeper with convolutions. In: IEEE Conf. on Computer Vision and Pattern Recognition (2015)
14. Teyssou, D., Leung, J.M., et al.: The InVID Plug-in: Web video verification on the browser. In: 1st Int. Workshop on Multimedia Verification (MuVer '17) at ACM Multimedia Conf. pp. 23–30. ACM (2017)