Multimodal Video-to-Video Linking

Turning to the Crowd for Insight and Evaluation

Eskevich, Maria; Larson, Martha; Aly, Robin; Sabetghadam, Serwah; Jones, Gareth J.F.; Ordelman, Roeland; Huet, Benoit

DOI
10.1007/978-3-319-51814-5_24

Publication date
2017

Document Version
Final published version

Published in
MultiMedia Modeling

Citation (APA)
Eskevich, M., Larson, M., Aly, R., Sabetghadam, S., Jones, G. J. F., Ordelman, R., & Huet, B. (2017). Multimodal Video-to-Video Linking: Turning to the Crowd for Insight and Evaluation. In L. Amsaleg, G. Þór Guðmundsson, C. Gurrin , B. Þór Jónsson , & S. Satoh (Eds.), *MultiMedia Modeling: 23rd International Conference, MMM 2017, proceedings* (Part II ed., pp. 280-292). (Lecture Notes in Computer Science; Vol. 10133). Springer. https://doi.org/10.1007/978-3-319-51814-5_24

Multimodal Video-to-Video Linking: Turning to the Crowd for Insight and Evaluation

Maria Eskevich[1(✉)], Martha Larson[1,2], Robin Aly[3], Serwah Sabetghadam[4], Gareth J.F. Jones[5], Roeland Ordelman[3], and Benoit Huet[6]

[1] CLS, Radboud University, Nijmegen, Netherlands
m.eskevich@let.ru.nl
[2] TU Delft, Delft, Netherlands
[3] University of Twente, Enschede, Netherlands
[4] TU Vienna, Vienna, Austria
[5] ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland
[6] EURECOM, Sophia Antipolis, France

Abstract. Video-to-video linking systems allow users to explore and exploit the content of a large-scale multimedia collection interactively and without the need to formulate specific queries. We present a short introduction to video-to-video linking (also called 'video hyperlinking'), and describe the latest edition of the Video Hyperlinking (LNK) task at TRECVid 2016. The emphasis of the LNK task in 2016 is on multimodality as used by videomakers to communicate their intended message. Crowdsourcing makes three critical contributions to the LNK task. First, it allows us to verify the multimodal nature of the anchors (queries) used in the task. Second, it enables us to evaluate the performance of video-to-video linking systems at large scale. Third, it gives us insights into how people understand the relevance relationship between two linked video segments. These insights are valuable since the relationship between video segments can manifest itself at different levels of abstraction.

Keywords: Crowdsourcing · Video-to-video linking · Link evaluation · Verbal-visual information

1 Introduction

Conventional multimedia information retrieval (MIR) research focuses on addressing an information finding scenario that starts with an ad hoc query, i.e., a query that is freely formulated by the user. The MIR system (or multimedia search engine) then returns results that are potentially relevant to the *information need* underlying the user's query, i.e., what the user had in mind to find. In this scenario the users have complete freedom in what they search for, since the MIR system is able to respond to any query. However, it has a significant disadvantage, if the users do not already have a clear goal in mind and are able to describe it, along with a reasonable understanding of the contents of the collection they are searching, then the 'freedom' to formulate their own query

L. Amsaleg et al. (Eds.): MMM 2017, Part II, LNCS 10133, pp. 280–292, 2017.
DOI: 10.1007/978-3-319-51814-5_24

is of little value. Effectively, the users run the risk of missing content that they might find relevant had they been shown it.

Video-to-video linking (also referred to as 'video hyperlinking') is a MIR scenario that has the potential to address the limitations of search based on ad hoc user queries. In the video-to-video setting, the users first enter a video collection by identifying an initially interesting video or video segment, but from there they can interactively explore the content in the collection by using video-to-video links to navigate between segments. Manual creation of all such links would be prohibitively time consuming, and thus automated methods of video-to-video linking are required. Such systems take an input video segment, which can be seen as analogous to a query, and which we refer to as an *anchor*, and then return a list of video segments, which we refer to as *targets*, to which it might be linked. A user exploring the collection will thus use a link to navigate from an anchor to a target.

A key challenge of video-to-video linking is to appropriately understand and model the relevance relationship between an anchor and its targets. The objective of a conventional MIR system is to seek items relevant to the information need behind the user query. Without a formal query, a video-to-video linking system lacks a well-defined information need. Further, without a way of characterizing the relevance relationship between anchors and targets, it is difficult to design video-to-video linking algorithms, and it is impossible to properly evaluate them.

To advance research in the area of video-to-video linking, we thus need to better understand this relevance relationship between anchors and targets that are automatically identified by a linking system. To address this issue in our work we adopt crowdsourcing methods. Consulting the crowd enables us to gather information from a large and diverse group representative of potential system users. This can also provide provide us with more insights into how users perceive video-to-video relevance. This paper describes our use of crowdsourcing in designing and evaluating the TRECVid 2016 Video Hyperlinking task [1], and the insights we gained on user interpretations of anchors.

The rest of the paper is structured as follows: in Sect. 2 we discuss the history of the video-to-video linking task. We then introduce an innovative new version of the task that focuses specifically on *verbal-visual information*: segments of video in which the videomaker is evidently exploiting, simultaneously, both the visual and the audio channel in order to communicate a message. In Sect. 3, we outline the specifics of the TRECVid 2016 Video Hyperlinking task, which is based on the concept of verbal-visual information. Section 4 explains our multistage crowdsourcing process for evaluating multimodal video-to-video linking systems. This combines an Anchor Verification stage, with evaluation of automatically identified targets. Then in Sect. 5, we discuss the results of the Anchor Verification carried out on Mechanical Turk (MTurk)[1]. In addition to verifying anchors, this task yielded insights about how people interpret video segments. Finally, in Sect. 6, we conclude and outline future work directions.

[1] www.mturk.com

2 Video-to-Video Linking

As identified in the Introduction, a key challenge in video-to-video linking is understanding and modeling the relevance relationship between two video segments in the absence of an existing information need formulated independently in the mind of the user. Rather we are seeking to determine whether to create a link on the basis that there might be such an information need in the future, when the user follows the link, although the user may follow the link for a more unfocused exploratory reason.

As a starting point we can observe that for two given video segments, it is possible to define an arguably infinite number of ways in which they can be related. For example, imagine two video segments of people sitting at a table: one from a video of a séance and the other from a talk show. The presence of the table in both segments allows us to assert that the segments are related, but this relationship is not necessarily interesting to users. In this case, the table is essential to the subject of neither video (both a séance and a talk show can happen without a table), nor does it serve to differentiate these segments from other segments. It is certainly possible that a user with great interest in tables would like to navigate from one table to another within a video collection. However, in focusing on providing paths through the collection for this particular user, we risk inadvertently inundating other users with too many possible relevance paths. In other words, rather then assisting them in locating information of interest in another video, we overload them with options to pursue.

It is important to note that comparing videos on the basis of detailed or exhaustive descriptions does not completely solve the problem. Continuing with our example of the video segments of the séance and the talk show: A detailed description would include information such as the names of the people talking, the transcript of what they are saying, a description of their clothing and appearance, and the details of the whole scene including the colour of the table, objects on it, and the background. If video segments are compared at this level of specificity, then we face a high chance that there are no related segments within the entire video collection. Although highly specific comparisons may exclude relationships between video segments exclusively based on the occurrence of incidental elements such as the table, they will also close off paths. The paths closed will be those that involve 'looser', somewhat unexpected relationships between anchor segments and target segments. It is exactly these links that connect video segments that are 'neither too loosely nor too tightly related' that will allow users to explore, and exploit, a video collection, without prior knowledge of what they will find in this collection.

2.1 Past Perspectives on Video-to-Video Linking

After work on automatically creating links in text [9,10] and automatically linking video anchors to text targets [8] and vice versa [2], at MediaEval 2012 we introduced the 'Search and Hyperlinking' (S&H) task [4]. The 'Search' part of the task was devoted to finding anchors in a collection of interest to users, which

then provided entry points for the 'Hyperlinking' part. The collection used was a large collection of semi-professional videos from blip.tv which represented a series of channels or 'shows' containing a series of episodes, adopting the style and structure of typical TV broadcasting (discussed further in Sect. 3.1).

S&H 2012 addressed the problem of making the links between anchor and target video segments neither 'too loose' nor 'too tight', by focusing the task on video segments that were likely to have interest to relatively many people in the user population. The anchors for S&H 2012 were selected by crowd workers, using an MTurk Human Intelligence Task (HIT) entitled, 'Find interesting things people say in videos'. Here, 'things' refers to the video segments that were used as anchors in the task. Workers were asked to consider the situation of sharing the video with another person, in order to nudge them to consider 'interesting' with respect to what other viewers would want to have in a video, and not exclusively with respect to their own personal interests.

Subsequently, in 2013–2014 the S&H task moved to professional broadcast content, and in 2015 became the TRECVid Video Hyperlinking (LNK) task. Now, in 2016, we return to both the blip.tv dataset, and also to the question of how to create links that are neither 'too loose' nor 'too tight' and also fairly evaluate them[2].

2.2 Verbal-Visual Information in Video-to-Video Linking

The new version of the video-to-video linking task was inspired by our realization that the 2012 instructions 'Find interesting things people say in videos' missed an important issue. People create video, and people watch video, not just because of what is *said* in video, but also because of what is *seen*. In other words, the intersection of the verbal and the visual modalities is the place to start looking for what users will perceive to be *important* about a video in situations where no independent information need can be assumed. In turn, this importance will provide the basis for solid judgments of video-to-video relevance.

A seemingly trivial observation that we can make about the users who created the videos in the blip.tv collection is that they chose to make videos and not audio podcasts (audio only) or to create images or silent video (video only). This observation is more important than it initially appears since it supports the assumption that people who make video choose video above other media because of the opportunity to exploit *multimodality*, i.e., both the audio and the visual channels. Further, it suggests that messages that videomakers find important to communicate to their audiences, can be found by simultaneously considering both the audio and the visual channels. If the videomaker puts effort into communicating a message to viewers exploiting both multimedia channels, we expect that viewers will have reasonable agreement on what this message is. Their perspectives will not completely converge, but they will tend to have stable interpretations of the aspects of a video message that are most important when

[2] For all HITs details, see: https://github.com/meskevich/Crowdsourcing4Video2VideoHyperlinking/

assessing the relevance relationship between two video segments. On the basis of this line of reasoning, in 2016, we decided to focus the LNK task on segments in the video in which viewers perceive that the *intent* of the videomaker was to leverage both the audio and video modalities to bring across a message to the audience. We refer to such segments as containing *verbal-visual information*.

Intent is a goal or purpose that motivates action or behaviour. In our context, we are interested in the goal the videomaker was trying to achieve in creating the video. The motivation for considering videomaker intent to be important derives from the investigation of uploader intent on YouTube that was carried out in [6]. Among the intent classes identified by the study are: 'convey knowledge', 'teach practice', and 'illustrate'. In [6], some initial evidence for a symmetry between the intent of users uploading video and the intent of users searching for (and viewing) video is uncovered. Our definition of verbal-visual information is agnostic to specific intent classes, such as 'convey knowledge', although addressing such classes might be of interest in future work. Here, our focus is on video segments in which the videomaker is presumably intentionally using both the verbal and visual channel. The driving assumption is that people viewing such segments will have relatively high consensus about what the videomaker is attempting to get across. It is not necessary, or even desirable, that there is complete consensus: rather our goal is to constraint possible relationships between the anchor and the target segments that viewers generally agree on, thereby supporting meaningful evaluation of video-to-video linking systems.

For concreteness, we return to the example of the two video segments above. By focusing on the presumed intent of the videomaker to communicate verbal-visual information, we have moved away from considering the table alone as the basis for linking, which would make the relationship 'too loose'. The reason is that the table is isolated in the visual channel, and is not part of the verbal-visual information that it was the videomaker's intent to bring across. At the same time, we do not insist on links based on similarity of the detailed descriptions of the two video segments, which would make the relationship 'too tight'. Instead, we focus on what is communicated by the videomaker through the interaction of words and visual content. In this way, the visual-verbal information of the video can take the place of the information need in the conventional MIR scenario. It will not be a unique source of relevance, but it will be universal enough to serve a large variety of users, and stable enough to be judged.

3 TRECVid 2016 Video Hyperlinking Task (LNK)

Next we overview the TRECVid 2016 Video Hyperlinking Task (LNK) [1].

3.1 Blip10000 Collection

For LNK at TRECVid 2016, we use the Blip10000 dataset which consists of 14,838 semi-professionally created videos [11]. In addition to the original Blip10000 collection, we used a new set automatic speech recognition (ASR)

transcripts provided by LIMSI [7] which uses the latest version of their neural network acoustic models. Using these transcripts and also detected shot boundaries [5], we indexed the entire collection to facilitate the process of defining anchors.

3.2 Defining Verbal-Visual Anchors

The task required us to define a set of anchors with respect to which video-to-video linking systems will be evaluated and compared. Since the set could not include all verbal-visual video segments in the collection, we sought to identify a good-sized subset. Because it is impractical to find these verbal-visual segments by randomly jumping into the collection, we used a search heuristic. Specifically, we searched the collection for 'linguistic cues' in the ASR transcripts. We defined 'linguistic cues' as short phrases that people typically use to signal that something seen rather than said is important to their overall message. We compiled the list of cues by reflecting on what people say when they are showing something, and arrived at the following list: 'can see', 'seeing here', 'this looks','looks like', 'showing', and 'want to show'.

Two human assessors (multimedia researchers) who were familiar with the video collection defined anchor segments by searching the blip.tv collection for occurrences of the cues in the ASR transcripts. They checked a 5 min window around each cue occurrence for verbal-visual information. Cases in which the cue did not lead to video containing verbal-visual information (i.e., someone says 'I can see what you are saying', but nothing is actually being shown in the visual channel) were skipped. For the other cases, an anchor was defined, by making a reasonable choice of a shorter anchor segment (10–60 s) within the five-minute window. The assessors also skipped cases that were very similar to previously chosen anchors, in order to ensure anchors diversity.

For each defined anchor, assessors wrote a short description of the anchor. In developing this summary, assessors attempted to abstract away from literal description of the content of the video. For example, a description would be 'Videos of people explaining where and how they live.' rather than 'Videos of a young man walking through the hall of an apartment and then explaining the contents of the bathroom.' The assessors controlled the level of abstraction by leveraging their familiarity with the collection to write a description for which they found it likely that more than one video segment in the collection could potentially be considered relevant.

3.3 Development and Test Set Anchors: Audio vs. Verbal-Visual

This section provides more details of the anchor sets released for LNK. Two anchor sets were released: 28 development anchors and 94 test anchors. The development anchors were those originally created for use in S&H 2012 [4]. They were defined by a crowdsourcing task carried out on MTurk. The wording of the HIT that was used to collect the anchor was given above in Sect. 2.1. Recall that

these anchors were focused on what people 'say' in videos. The Anchor Verification stage of our evaluation process (see Sect. 4.1) determined that a number of these indeed involved verbal-visual information, although they were not created explicitly to do so. The verification stage assured us of the appropriateness of the 2012 anchor set for development purposes in 2016.

4 Crowdsourcing Evaluation

Our evaluation takes the form of three stages (Anchor Verification, Target Vetting, Video-to-Video Relevance Analysis), each realized as a HIT on MTurk. We discuss each in turn. The principles informing the design of our HITs are: (1) provide descriptions that avoid technical terms, but rather allow workers to identify with context of use, and, (2) user quality control mechanisms that are fair and also an integral part of the HIT.

4.1 Stage 1: Anchor Verification

The Anchor Verification stage verifies the verbal-visual nature of the anchors defined by the assessors. Specifically, it checks whether the perceptions of viewers (i.e., potential users of the video-to-video linking system represented by workers on MTurk) align with those of the assessors. We related the HIT to a context of use with which we hope the crowd workers can identify by entitling it 'Watch the video segment and describe why would someone share it'. Further, we avoided reference to 'videomaker intent' or 'verbal-visual information' and instead provided the workers with the following instructions. *We know that people upload those videos for certain reasons. We ask you to think about what the person who made the video was trying to communicate to viewers during this short piece. In other words, what are viewers supposed to understand by watching this particular short piece of the video.* The HIT then asked them to provide a description of the anchor. We asked 'How would you describe the content you see to another person?' We collected a description from three different workers for each video. We then went through the descriptions that the workers provided, and checked whether they contain reference to an element present in the visual channel of the video. The anchors that were not verified to be of verbal-visual nature were not used in the subsequent crowdsourcing stages.

In designing the HIT, we also tried to guide the workers away from highly specific descriptions, since we are not interested in 'tight' relationships between anchors and targets, as discussed above in Sect. 3.2. We used two mechanisms to encourage workers to be more abstract. First, we included three example descriptions (two taken from the originally collected development anchor set and one from the test anchor set). These examples were intended to convey to the workers the diversity of the videos in the collection, and the level of abstraction of description that we were targeting. The examples included anchors bearing the descriptions 'We are looking for videos of people explaining where and how they live.' and 'This video is about features of a computer software application that

supports communication.' Second, we stated, 'Please keep in mind that the this description that you write will be shown to someone else (working on Amazon Mechanical Turk), who will be asked to find other videos that are related to this video segment.' As discussed further in Sect. 5, the workers' answers gave us insight into how people abstract from literal descriptions of video clips, to more abstract descriptions.

4.2 Stage 2: Target Vetting

Target Vetting is the first step in assessing the output of video-to-video linking systems. The systems participating in the benchmark returned a list of targets, i.e., video segments, for each anchor in the test set. Target Vetting compares potential targets to the textual description of the anchors originally generated by the assessors. We use Target Vetting because the potential number of targets is very large. A crowdsourcing task to make a video-to-video judgment between anchor and target for every generated target would be prohibitively large. Further, the density of relevant anchor/target pairs could be low, making the task unfulfilling, tedious, and potentially impacting worker performance.

The output of the Target Vetting task is a set of binary decisions about the relevance relationship between the anchor description, and each target that each system has generated for that anchor. We do not ask workers to make the decision directly, since it is hard to enforce consistency in the criteria that they use to make the comparison. Rather we used multiple choice questions which are a good way of creating consistency, since the list of choices conveys information about the answer space and the types of distinctions that are relevant for answering the question. For this reason, the Target Vetting HIT asked workers to watch the target video segment, and then to choose among five anchor descriptions. If the target is relevant to the anchor, the worker should be able to pick the correct anchor description. We force the workers to choose between the five answers, and then ask them whether they are happy with the choice that they made, and whether it was easy to make the decision. If the target is not relevant to the anchor, the worker should find it difficult to pick a correct anchor description, and should express dissatisfaction or discomfort with the pick. Note that asking about the level of discomfort plays the same role as including a 'none of the above' option. The 'forced-choice' question that we use is, however, superior to a 'none of the above' option since the workers need to judge their own reaction (which is familiar territory) rather than contemplate whether or not there might exist an answer that is more fitting than any in the list (which is open ended). The four out of five descriptions are chosen randomly from the dataset, and sometimes could fit the target better than the ones of the original anchors. These cases might help to extend the ground truth for the other anchors, while the current anchor in question earns a non-relevance judgment.

4.3 Stage 3: Video-to-Video Relevance Analysis

Video-to-Video Relevance Analysis is the second step in assessing the output of video-to-video linking systems. Because targets that were not potentially relevant to anchors were eliminated in the previous stage, the crowdsourcing task at this stage can be smaller, and its output can be manually analyzed. This stage consists of a HIT that presents workers with two video segments, the anchor and the corresponding target, and asks them to describe the way in which the two are related. We do not provide choices or suggestions about how the anchor and target could be related (i.e., we do not suggest 'person'/'object'/'speech'), or otherwise constrain answers. Instead, we collected unstructured information in the form of 2–3 natural language sentences. The first stage, Anchor Verification, checked that we were focused on verbal-visual information. We believe that the Video-to-Video Relevance Analysis stage gives us comprehensive qualitative information about the types of relevance between video segments that people perceive. We hope to find relevance aspects that we could not have envisaged beforehand, and also to better understand diversity in relevance relationships, in particular, with respect to the different levels of abstraction at which relevance relationships are perceived.

5 Insights from the Crowd

In this section, we dive more deeply into the insights that are gained from Stage 1: Anchor Verification, described in Sect. 4.1. We ran the HIT on MTurk, with a 0.15 USD reward and restricted to workers with overall 90% HIT Approval Rate. We ran the HIT on 119 anchors (we reduced the original set of 28 development and 94 test sets by taking three anchors as examples) and collected 357 worker judgments. The quality of submitted work confirmed that our HIT design was clear and easy to grasp. There were only two cases in which the submissions had to be rejected for unserious work. A number of workers left comments with thanks, and one reported having enjoyed the task. For each anchor, we compared the anchor description originally provided by the assessors with the workers descriptions both individually, and as a set of three (since three responses were collected per anchor). In total, 51 unique workers carried out the task, with an average of seven judgments per worker, with a maximum of 59 HITs in case of one worker. The rest of this section reports our findings.

5.1 Verbal-Visual Information

The primary purpose of Stage 1: Anchor Verification was to eliminate anchors that were not truly verbal-visual in nature from the test set. Table 1 provides an example of a case where we judged that none of the three workers' descriptions mentioned a visual element to the anchor. As mentioned in Sect. 4.1, such anchors were dropped. Three of the original anchor test set defined by the annotators were affected, leaving us with a final test set of 90 anchors.

We also used Stage 1 to verify whether any of the anchors in the development set could be considered to contain verbal-visual information, although they were

Table 1. Example of an anchor that failed crowd Anchor Verification: it did not contain the visual component necessary to be verbal-visual information, and was dropped.

Anchor ID	Original assessor description	Description by MTurk worker
89	We are looking for videos of a modern, enthusiastic preacher, pastor or spiritual leader	In this video a speaker is talking about how lack of movement and lack of socialization in a social media and computer generation is causing obesity
		A young man is trying to convey to his audience that too much time spent on the computer is leading to bad health and that people need to spend more time meeting each other face to face rather than with social media online
		The video explains how the overuse of social networking sites cause health problems. This is a very important point to note

created by asking about spoken content. It turned out to be the case that some development anchors were indeed verbal-visual. An example of such a case is Anchor ID 25, whose description is 'The launch of a new villain for the Avengers to fight with.' The worker descriptions suggest the critical contribution of an image of Red Skull to the message conveyed by the anchor.

5.2 Observations on Abstraction

Stage 1: Anchor Verification also yielded some insights on the level of abstraction at which people describe video. Recall that with our HIT design, which included examples, we tried to encourage the workers to provide us with abstract anchor descriptions. In general, however, the workers descriptions were much more specific than the descriptions provided by the assessors who originally defined the anchors. Table 2 illustrates such cases. In a given collection, it would be difficult to find two video segments that both fit this description, unless the collection had a special nature or the video segments were near duplicates. We did, however, find cases in which the workers descriptions reached the same level of generality or abstractness as the original descriptions. Table 3 illustrates two examples.

Finally we would like to mention that when analyzing the worker description, we discovered that workers make reference to what we call the *conventional purpose* of the video. We define conventional purpose categories as the categories into which people conceptualize videos. Any category that is widely acknowledged, i.e., conventionally accepted, can be considered. However, we point out the our use of the word *purpose* is motivated by the fact that these categories involve typical situations in which videos are watched and reasons why people watch them. For example, some workers made comments about the genre (that it

Table 2. Examples of anchors for which workers provided detailed descriptions of the literal video content.

Anchor ID	Original assessor description	Description by MTurk worker
106	We are looking for videos that show a collection of people's opinions on the street	This video shows that there is a question as to whether there is a youth movement for voting. There are quick interviews with people in their late teens or early 20 s on the topic
74	We are looking for videos of someone fumbling or dropping something by accident	The man on the left is clearly someone of higher status who is ordering his butler (or similar position) to give him his phone. The butler (man on the right) points out that the phone is in the man's hand, and refuses to pick it up when the man clumsily drops it on the floor

Table 3. Examples of anchors where the level of abstraction is matched between the original description and the description provided by the worker.

Anchor ID	Original assessor description	Description by MTurk worker
77	We are looking for videos that give ideas of different kinds of purses and handbags	This video goes over some of the different variations in women's handbags
107	We are looking for videos with tips about using social networking sites such as Facebook	Video about posting to social media. Showing how to physically post information

was a news report or a comedy) and others pointed to other categories reflecting what the videos are used for (that it is for selling something). We observed them mentioning in the description that the video was part of a lecture, commercial, or blooper reel, and that the video was a performance. These references to what the video is used for are also a form of abstraction. We believe that ultimately these sorts of purposes are related to the intent categories mentioned in Sect. 2.2, and are eager to explore user use of these categories further.

6 Summary and Outlook

This paper provided a brief review of the history of video-to-video linking, and introduced a new multimodal video-to-video linking scenario. This new scenario exploits the assumption that users will have more stable judgments about the relevance relationship between anchors and targets, if we focus on cases in which

the videomaker is intentionally using both audio and visual information stream to convey a message. We introduced a crowdsourcing strategy for evaluating systems that address this scenario, and also presented findings from the first stage of this evaluation 'Anchor Verification'.

Our next step is carrying out Stage 2 and Stage 3 of the evaluation of the 2016 benchmark, once the participants have submitted their system output. We point out that we anticipate on the basis of general human behaviour that we will find a mismatch between generation and validation: we expect that although workers tend to generate descriptions at a literal level, that they will not insist on this level when they are choosing a description that fits a video segment (i.e., in Stage 2: Target Vetting). Further, we expect to discover that workers will also more naturally move to higher levels of abstraction when comparing anchors and targets (i.e., in Stage 3: Video-to-Video Relevance Analysis).

Additionally, we would like to pursue further insights already revealed by Stage 1: Anchor Verification. We are particularly interested in the relationship between the level of abstractness at which people describe video segments, and their knowledge of the collection from which the video segments are drawn. Here we have seen that the assessors who are familiar with the collection describe videos with a higher level of abstraction. If collection familiarity impacts that level of abstraction that is most appropriate for the video-to-video relevance relationship in the linking scenario, the implications for systems that attempt to hyperlink large collections could be profound.

Acknowledgments. This work has been partially supported by: ESF Research Networking Programme ELIAS (Serwah Sabetghadam, Maria Eskevich); the EU FP7 CrowdRec project (610594); BpiFrance within the NexGenTV project, grant no. F1504054U; Science Foundation Ireland (SFI) as a part of the ADAPT Centre at DCU (13/RC/2106); EC FP7 project FP7-ICT 269980 (AXES); Dutch National Research Programme COMMIT/.

References

1. Awad, G., Fiscus, J., Michel, M., Joy, D., Kraaij, W., Smeaton, A.F., Quénot, G., Eskevich, M., Aly, R., Jones, G.J.F., Ordelman, R., Huet, B., Larson, M.: TRECVID 2016: Evaluating video search, video event detection, localization, and hyperlinking. In: Proceedings of TRECVID 2016, NIST, USA (2016)
2. Bron, M., Huurnink, B., Rijke, M.: Linking archives using document enrichment and term selection. In: Gradmann, S., Borri, F., Meghini, C., Schuldt, H. (eds.) TPDL 2011. LNCS, vol. 6966, pp. 360–371. Springer, Heidelberg (2011). doi:10.1007/978-3-642-24469-8_37
3. Eskevich, M., Jones, G.J.F., Larson, M., Ordelman, R.: Creating a data collection for evaluating rich speech retrieval. In: Eighth International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, pp. 1736–1743 (2012)
4. Eskevich, M., Jones, G.J.F., Chen, S., Aly, R., Ordelman, R.J.F., Larson, M.: Search and hyperlinking task at mediaeval 2012. In: MediaEval CEUR Workshop Proceedings, vol. 927, CEUR-WS.org (2012)

5. Kelm, P., Schmiedeke, S., Sikora, T.: Feature-based video key frame extraction for low quality video sequences. In: 10th Workshop on Image Analysis for Multimedia Interactive Services (2009)
6. Kofler, C., Larson, M., Hanjalic, A.: User intent in multimedia search: a survey of the state of the art and future challenges. ACM Comput. Surv. **49**(2), 1–37 (2016)
7. Lamel, L.: Multilingual speech processing activities in Quaero: application to multimedia search in unstructured data. In: The Fifth International Conference Human Language Technologies - The Baltic Perspective Tartu, Estonia, 4–5 October 2012
8. Larson, M., Newman, E., Jones, G.J.F.: Overview of videoCLEF 2009: new perspectives on speech-based multimedia content enrichment. In: Proceedings of the 10th International Conference on Cross-language Evaluation Forum: Multimedia Experiments (CLEF 2009), Corfu, Greece, pp. 354–368 (2009)
9. Mihalcea, R., Csomai, A.: Wikify!: Linking documents to encyclopedic knowledge. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management (CIKM 2007), Lisbon, Portugal, pp. 233–242 (2007)
10. Milne, D., Witten, I.H.: Learning to link with Wikipedia. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM 2008), Napa Valley, California, USA, pp. 509–518 (2008)
11. Schmiedeke, S., Xu, P., Ferrané, I., Eskevich, M., Kofler, C., Larson, M., Estève, Y., Lamel, L., Jones, G.J.F., Sikora, T.: Blip10000: a social video dataset containing SPUG content for tagging and retrieval. In: Dataset Track. ACM Multimedia Systems, Oslo, Norway (2013)