Creating HAVIC: Heterogeneous Audio Visual Internet Collection

Stephanie Strassel[1], Amanda Morris[1], Jonathan Fiscus[2], Christopher Caruso[1], Haejoong Lee[1], Paul Over[2], James Fiumara[1], Barbara Shaw[2], Brian Antonishek[2], Martial Michel[3]

1 – Linguistic Data Consortium (LDC), 3600 Market Street, Suite 810, Philadelphia, PA 19104

2 – National Institute of Standards and Technology (NIST) Technology, 100 Bureau Dr., Gaithersburg, MD 20899

3 –Systems Plus, Inc., One Research Court -Suite 360, Rockville, MD 20850

E-mail: strassel@ldc.upenn.edu, amandamo@ldc.upenn.edu, jonathan.fiscus@nist.gov, carusocr@ldc.upenn.edu, haejoong@ldc.upenn.edu, paul.over@nist.gov, jfiumara@ldc.upenn.edu, barbara.shaw@nist.gov, brian.antonishek@nist.gov, martial.michel@nist.gov

Abstract

Linguistic Data Consortium and the National Institute of Standards and Technology are collaborating to create a large, heterogeneous annotated multimodal corpus to support research in multimodal event detection and related technologies. The HAVIC (Heterogeneous Audio Visual Internet Collection) Corpus will ultimately consist of several thousands of hours of unconstrained user-generated multimedia content. HAVIC has been designed with an eye toward providing increased challenges for both acoustic and video processing technologies, focusing on multi-dimensional variation inherent in user-generated multimedia content. To date the HAVIC corpus has been used to support the NIST 2010 and 2011 TRECVID Multimedia Event Detection (MED) Evaluations. Portions of the corpus are expected to be released in LDC's catalog in the coming year, with the remaining segments being published over time after their use in the ongoing MED evaluations.

Keywords: Multimodal, Event Detection, Annotation, Data Centers, Data Collection and Distribution

1. Introduction

The lifeblood of current video and language technology development is data. Data enables everything from algorithm development to the generation of new ideas. To support new research directions in multimodal event detection and related technologies, Linguistic Data Consortium (LDC) in collaboration with the NIST[1] Multimodal Information Group is developing a large, heterogeneous annotated multimodal corpus. The Heterogeneous Audio Visual Internet Collection (HAVIC) will ultimately consist of thousands of hours of unconstrained multimodal data, annotated for a variety of features. The primary focus for the HAVIC corpus is user-generated videos with content occurring in the audio, video, and text embedded in the video.

Progress toward developing new technologies is typically driven through use of constrained, technology-focused, controlled data sets. An example of this approach is the NIST Speech-To-Text (STT) evaluation series (NIST 2012), in which data sets have been designed to be neither too difficult nor too easy for current technology capabilities. As technologies improve, new domain challenges are introduced to the data, removing earlier constraints and driving continued technology improvements. While focused-domain data sets may enable technology progress in one modality, the chosen domain can lack the challenges necessary to promote progress across different modalities. For instance, the Meeting Domain presents many acoustic challenges in the speech modality, making it a good choice for driving fundamental STT research. The same domain however lacks variability in the video modality, making it less interesting to most video researchers.

The HAVIC corpus has been designed with an eye toward providing increased challenges for both acoustic and video processing technologies. Video and audio technologies are maturing to the point where a domain that includes challenges for multiple modalities will foster multimodal research. HAVIC was designed with such challenges in mind, targeting the multi-dimensional variation inherent in user-generated video content, including variable camera motion, subject topicality, low and high quality video resolution and compression, competing background noise, spontaneous or concurrent speech, far field speech, multiple languages, and so on.

2. Multimedia Event Detection

The Multimedia Event Detection (MED) task, to be run in TRECVID for several years, represents the first application of the HAVIC collection to the evaluation of multimedia systems. The goal of MED is to assemble core detection technologies into a system that can quickly and accurately search a multimedia collection for user-defined events that include a person interacting with another person or object. Events like making a cake or assembling a shelter are defined in advance, and a portion of the HAVIC corpus is comprised of videos that illustrate these events. The corpus also includes negative examples (i.e. videos that may be superficially similar to the events but fail to satisfy the event definition) and a large number of videos that are completely unrelated to

[1] *Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.*

the events. Each video added to the corpus is also manually annotated with a set of judgments describing its event properties and other salient features.

2.1 Event Selection and Definition

The first stage of building the HAVIC corpus is event development. LDC created a pool of 164 candidate events from which NIST selected a final set of 75. These 75 events were then divided into 7 "event sets", typically consisting of 10 to 15 events each. The events were selected according to a rigorous process to both protect privacy and appropriately sample the domain space so that the evaluations against it would provide a meaningful measure of research progress along several dimensions. The details of the selection process and its results cannot be made public until after the conclusion of the MED evaluations to protect the blindness of the evaluations.

2.2 Event Kits

For each of the 75 official events selected, LDC then creates a textual "event kit", which is a textual description of the event properties along with a few exemplar videos. The event kit is used during manual annotation of collected HAVIC videos, as well as in the MED evaluation. Each event kit consists of

- An **event name** which is a mnemonic title for the event.
- An **event definition** which is a textual definition of the event.
- An **event explication** which is a textual exposition of the terms and concepts used in the event definition, at least those not commonly known.
- An **evidential description** is a textual listing of attributes that are indicative of an event instance. The evidential description provides a notion of some potential types of visual and acoustic evidence that might be indicative of the event having occurred in a given video. The evidential description is not intended to be an exhaustive list of attributes, nor should it be considered a list of required evidence.
- A set of **exemplars**, which are illustrative video examples each containing an instance of the event. The examples are illustrative in the sense that they help form the definition of the event but they may not demonstrate all possible variability or potential realizations of the event.

Figure 1 contains the event kit text for the "Getting a vehicle unstuck" event used in the 2011 MED evaluation.

3. Collection and Annotation

Once events have been defined, data collection and annotation begins. LDC oversees a large team of human annotators known as data scouts to search for suitable video content and annotate videos for a variety of features. While there is a potentially unlimited supply of user-generated video content on the Internet, not all of it is suitable for inclusion in HAVIC and an important part of the data scout's job is discerning what videos should be excluded. One part of this consideration is the status of a given video with respect to intellectual property rights (IPR). HAVIC data presents particular challenges in the domains of copyright and contract law, privacy and objectionable content. Although video clips may originate from anywhere in the world, HAVIC's conservative assumption is that all video clips are copyrighted, and LDC takes additional steps to ensure that collected data can be redistributed for research, education and technology development under the Doctrine of Fair Use. As such, scouts are given a list "approved" video host sites whose terms of use are compatible with the intended use of the HAVIC corpus. Further, scouts are instructed to select individual videos with an appropriate license, for instance a type of Creative Commons license that permits redistribution.

Event name: Getting a vehicle unstuck
Definition: One or more people work to free a motorized or unmotorized vehicle that is stuck.
Explication: A stuck vehicle is one that either cannot move, or can only move in a limited area (e.g. a parking space). Usually movement is restricted either because the vehicle is located on/in a substance that prevents the wheels of the vehicle from having enough friction to propel the vehicle forward (e.g. mud, mud covered in water, snow, etc.), or because the angle of the vehicle and/or surrounding obstacles prevent it from moving normally. Getting a vehicle unstuck can be done by maneuvering the vehicle itself or by using another device (another vehicle, an object, etc.) with the intention of getting the vehicle to a location or angle where it can again move normally.

Evidential description
scene: typically outdoors, day or night, any weather
objects/people: vehicle (car, boat, bicycle, construction vehicle), person operating the vehicle, possibly other people assisting in or outside of the vehicle, possibly additional vehicle(s) assisting
activities: steering, hitching, chaining, driving, sliding, skidding, pulling
audio: vehicle tires spinning against surface; engine noise from vehicle; narration or commentary on process from participants.

Figure 1: Event Kit for Getting a Vehicle Unstuck

The scouts were also required to identify and exclude clips likely to contain commercially-produced copyrighted content. To further protect the privacy of data creators and to ensure that the corpus does not contain problematic content, all videos are manually screened prior to inclusion in the corpus. The video itself

must be free of inappropriate content including sensitive personal identifying information (PII) or sensitive content (illicit activity, extreme profanity, nudity, etc).

3.1 AScout Framework and the Annotation Process

In order to provide a structured framework for data scouting and annotation, and to minimize the amount of aimless web surfing, LDC has developed a web-based user interface and backend framework for HAVIC collection and annotation, known as AScout. The AScout user interface is a Firefox add-on consisting of an annotation form displayed on the left side of the browser window. Data scouts use the browser in the usual way to search, navigate video websites and watch videos. When they find a suitable video they fill out the AScout form on the left side of the screen, and the results are logged to a database.

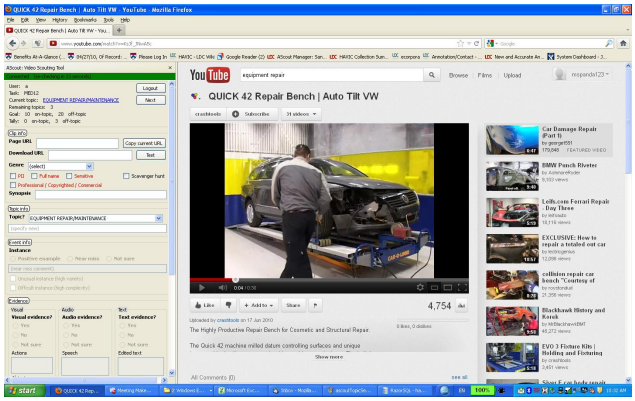

Figure 2: AScout Annotation Interface

Upon signing in to AScout, data scouts are given a specific scouting assignment, which typically consists of a target number of instances for a specific event plus some number of "background clips" (i.e., videos which do not contain one of the MED events). For instance, in a given scouting session a scout may be directed to find 10 "making a cake" videos and 25 "background" videos, and a counter in the AScout interface shows progress toward that goal.

AScout requires annotators to make a number of decisions about each video they submit. The first is the page URL for the clip, along with the likely download URL.

After logging the video URL, the data scout assigns the video to a genre (e.g. home movie, how-to) and writes a brief synopsis of the video content. Scouts also flag any problematic content, such as PII. If a video is flagged at this stage, no further annotation occurs and the video is excluded from the HAVIC corpus.

For videos without problematic content, the next annotation task is to assign a topic category, using a pull-down list in AScout. There are two types of topics in AScout: general topic categories like FOOD & DRINK or SPORTS; and event-based topics like LANDING A FISH or MAKING A CAKE that correspond to an official MED event. If a video clip being submitted meets the definition of one of the MED events, annotators choose that event-based topic from the pull-down list. Otherwise, they select the general topic that is most relevant to the video.

Background video clips require no further annotation beyond noting the license type for the video. Clips that contain a MED target event require several additional annotations. First, the data scout must determine whether the video shows a true instance of the event (a positive instance), or whether the video does not quite meet the requirements for a positive instance (a near miss). To make this determination, the data scout consults information in the textual event kit, such as the event name, definition, and explication.

Beyond the event kit, annotators follow two guiding principles to help determine whether a video is a positive instance or not. First, a **Sufficient Evidence Rule** states that the video must contain sufficient evidence that the event has occurred in order for the annotator to call it a positive instance. Evidence can come from any part of the video: visual, audio and/or text embedded in the video. A corollary to this rule is that it is not necessary for a clip to contain *every* part of the event process in order for the annotator to conclude that the event has occurred. The data scout must also follow a **Reasonable Viewer Rule**, which states that if according to a reasonable interpretation of the video the event must have occurred, then the clip is a positive instance of the event.

If the data scout concludes that the video does indeed contain an instance of a MED event, the last type of annotation performed is to summarize the type of evidence that led to this conclusion. First, the data scout indicates the presence or absence of visual evidence, audio evidence, and text evidence. Scouts also have the option of listing individual items that they considered to be evidence for the event; for instance they can enumerate the particular objects, people, or activities shown in the video, or particular words spoken or shown in text. They can also flag the presence of foreign-language speech or text, and indicate if steps of the event are narrated in the video.

3.2 Harvesting and Processing

When a data scout submits the AScout annotation form, it is sent to a server at LDC that stores the annotation (including the video URL) as a new row in the HAVIC annotation database, with a sequential identification number. In order to reduce the likelihood of duplicate video submissions, the AScout annotation framework includes a URL-based duplication checking routine. For many video host sites, a single video could be accessed via several different URLs, depending on when the video was accessed or how the viewer searched for it. To address this, the AScout framework includes a URL normalization function that standardizes variant URLs for the same video on a host site, thus increasing the chances that a duplicate URL will be detected. The md5 hash of the normalized URL is stored with the

annotations in the database. When the data scout enters the video URL into the AScout form, AScout normalizes it and checks the database to see if its md5 hash already exists there. If so, AScout warns the user and disables submission of the form. Despite these measures some amount of duplicated content is unavoidable since users may post or re-post the same video, or portions of the same video, at different times or on different host sites.

Harvesting submitted videos is the next piece of the process. The annotation database is checked by site-specific instances of a downloader script at regular intervals, ranging from once every five minutes to once daily. Sites which contain session timeout tokens in their URL string require frequent checks, but not so frequently that the downloader interferes with website traffic or causes redirection to a CAPTCHA test (which would result in an unsuccessful download).

Each video host site requires a specific set of database queries and download execution syntax, which are parsed in from XML configuration files. After this information is read, the downloader script queries the database and generates a list of URLs. It also retrieves the annotation's unique sequential ID number, which is used as the output file root name.

During the download step, the script first checks the clip repository to see if the file already exists. If not, it executes the clip download command - typically wget or curl, but frequently an open-source application unique to a single site – and then performs a clip check and conversion routine once the download has been completed.

Media files in the corpus are required to be in MPEG-4 format, with h.264 video encoding and AAC audio encoding. The clip checker/converter verifies that the file is a valid video file. If it is already a valid video file in the correct format, then the database is automatically updated with the video file name and the script moves on to the metadata generation step. If the file is a different video format, such as Flash, Ogg, Matroska, or a non-h.264 MPEG, the video is automatically converted to the required format using either ffmpeg or mencoder depending on the input format. Original video resolution and audio/video bitrates are retained, which results in a wide variety of frame rate, resolution, and bit rate in order to reflect data as found in the wild. If the file is not a valid video format, such as in the case where a video is removed and the downloader retrieves a 404 html file, the annotated clip is flagged for a retry during the next download cycle. If it fails again, the annotation is listed as a failed download.

After the checks and conversion, additional clip metadata is generated including video duration, codec, unique randomized ID number, and md5 checksum.

3.3 Annotation Quality Control

The volume requirements on the HAVIC corpus are extremely demanding. To support the evaluation goals of MED, the HAVIC corpus will eventually contain hundreds of thousands of videos. To create such a large corpus on a fixed timeline requires a large team of data scouts and several managers to oversee and check their work. The current team of over 60 data scouts annotates hundreds of videos each day.

To be meet the volume and quality requirements of the HAVIC corpus, data scouts must be efficient, accurate and creative. Candidate scouts are required to take a pre-screening test that assesses their ability to quickly locate novel video content on the web under time pressure, and to identify problematic content (which should not be included in the corpus) in a set of existing videos. Scouts who pass the preliminary screening undergo a roughly two week training process to familiarize them with the goals of the HAVIC collection, event definitions and other project requirements.

Ongoing data scout training includes biweekly group meetings, review of individual annotators' work by senior project staff and other quality control measures. For instance, a portion of all videos in the corpus are reviewed during a second pass by senior annotators. Event instances are highest priority for second passing, but a random sampling of background clips are also reviewed. The primary goal of second passing is to make sure the video's annotations are accurate and also that the video is usable. Additional quality control is performed continuously via manual inspection of the annotation table. LDC further maintains a mailing list and wiki for data scouts to discuss questions, share scouting tips and record decisions.

The nature of the HAVIC data scouting task lends itself to annotator variation on a number of dimensions. One of these is the assignment of a topic to each video. The pre-defined list of general topics are broad categories, not meant to perfectly partition the world of videos, and so there will often be videos that could fit in multiple topic categories. In the case where a video fits multiple general topics, data scouts are instructed to select the topic they find most appropriate. A video showing a child walking a dog might be given the topic CHILDREN by one annotator, ANIMALS by another annotator, and OUTDOOR ACTIVITIES by a third annotator. This is an expected feature of the corpus.

Another source of annotation variation relates to the event-specific topics and the fact that the corpus is not exhaustively annotated for every event. When a data scout observes an instance of a MED event they are instructed to label it as such, but it is not expected that scouts will consider every clip they view against the entire set of MED events. In fact, because the MED events were developed over time, not all events were known during early phases of the corpus collection and annotation process. As a result, the corpus is known to contain some number of videos that are assigned to general topics but which depict a MED event. Also, some videos may contain instances of more than one MED event without being explicitly labeled as such. When it comes to assessing individual videos against a single topic, annotators typically show high agreement.

That is to say, most of the time most annotators agree about whether a given video does or does not contain a positive instance of an event like MAKING A SANDWICH. However, we expect there to be a number of judgment calls where well-trained, reasonable annotators simply disagree with one another. For instance, if the video shows a person spreading butter on a piece of bread, folding it in half and eating it, annotators may not agree on whether that video is a positive instance of MAKING A SANDWICH. This type of annotator variation is expected, and accepted, in the corpus though it has not been quantified or measured to date.

Because one goal of the HAVIC corpus is to reflect the challenges of real-world multimedia data, corpus variety is key. Data scouts are encouraged to seek out variety along almost any parameter. To encourage creative scouting LDC uses a variety of techniques that may include games and contests. One such game is the scavenger hunt, where data scouts are given a specific theme and asked to conduct searches with that theme in mind, with incentives awarded for creative interpretations of the theme. Scavenger hunts may focus on the abstract ("almost") or the concrete ("videos involving more than 10 people"), and are intended to increase the variety of the corpus as a whole but also to ensure that the set of collected positive instances of each event reflect the whole spectrum of realizations of that event in real-world data. To further support this goal LDC utilizes resources like WordNet (Princeton 2010) to identify terms and concepts that could yield additional search strategies.

4. Corpus Distribution

To date, portions of the HAVIC corpus have been distributed to performers in the TRECVID MED evaluation task. For each MED evaluation NIST selects a portion of the corpus for use as training, development test or evaluation test and LDC distributes the data to performers, with video data shipped out on hard drives, and corresponding documentation and metadata distributed as a web download package. Each video package consists of the video files in mp4 format with h.264 video encoding and aac audio encoding. Video resolution and audio/video bitrates are retained as found in the original harvested files, and md5 checksums, clip duration and codec information is provided. The documentation and metadata packages contain information including the EventIDs and Event Names for the event set(s) targeted in this release as well as an event metadata table showing (partial or complete) annotations for some clips.

To date the HAVIC corpus has been used to support the NIST 2010 and 2011 TRECVID Multimedia Event Detection Evaluations. Portions of the corpus are expected to be released in LDC's catalog in the coming year, with the remaining segments being published over time after their use in the ongoing MED evaluations.

5. References

NIST "Rich Transcription Evaluation Project" Website. 2012. http://www.itl.nist.gov/iad/mig/tests/rt/

Princeton University "About WordNet." WordNet. Princeton University. 2010. http://wordnet.princeton.edu