Dr. Sibt ul Hussain Rafia Rahim Talha Hanif Butt

Vid2Info: A video analysis system for TV streams

A fully automated system for recording, processing, storing, indexing and analyzing the contents of TV programs.

Introduction: What it is for?

This document discusses the feasibility of a fully automatic end-to-end solution for the compression, recording, processing and archiving of up to 250 channels of video (TV footage). The proposed system will provide the transcribing, tagging and analysing of aired TV content in near real time. In short, it will serve as search engine for the braodcasted TV channels.

There is a huge amount of video data that is available in the form of videos from TV channels and this amount of data is increasing at exponential rate. However, unfortunately no such system exists – according to our best of knowledge – that can automatically and reliably transcribe the Urdu channels. This is quite unfortunate, given the amount of progress machine learning in general and computer vision in specific has seen in recent years. Our first major goal here is to exploit the recent advances in machine learning, deep learning, and computer vision and to build a system that can automatically and reliably transcribe the content and extract all the relevant information present in the transmitted videos. Secondly, we want to build quality data products such as news sentiments, obscene content detection and filtering, *etc.*, from the extracted information, for more information please refer to detailed description of the features.

Existing Solutions?

There are multiple existing commercial solutions that provide either only crude storage and processing facilities or end-to-end solutions with indexing and searching of content.

Crude Storage & Processing Facilities: These services can be further subdivided based on their infrastructure type, i.e. cloud-based or local-server based. For cloud based services, Huawei Video Cloud Storage Solution[1] (Google Cloud, AWS S3 can be considered but are costly alternatives) seems a reasonable option. As it supports the collection and storage of up to 384 HD video-stream channels, with up to 96 channels of

1 http://e.huawei.com/us/solutions/business-needs/data-center/storage/video-storage

simultaneous play-back. It is a fully configured system that can collect and store hundreds of thousands of videos concurrently. However, this system is a very crude system and does not provide any automatic in-depth tagging and indexing of data based on its content. Secondly it only gives video analysis based on just 30 previous days content and still misses many important features that can help in more detailed analysis.

A more viable alternative will be to utilize the off-the-shelf video storage solutions, *e.g.*, OceanStor 2800 V3 Video Cloud[2], to build local server based setup. One OceanStor 2800 V3 device supports concurrent recording of 400 HD videos and concurrent playback of 100 HD videos with the possibility of connecting of six physical servers. Thus OceanStor can act as storage server and rest of the features can be build on top of this server.

End to End Solutions: There exists solutions that provide end-to-end video storage and retrieval from video streams. Some of the examples in this category include: voiceinteraction[3], tveyes[4], cyberalert, cision, *etc.*, still none of these systems provide the required and requested features for basic and advanced analysis. As majority of them are targeted to US or Europe market and thus do not have support for our local channels. Furthermore, building customizable solutions on top of them is extremely hard due: (i) to close-nature of these softwares; (ii) to absence of uniform APIs; and (iii) to their prohibitive costs.

Thus there is a need to build an end-to-end system using available open source components that can serve the required purpose and provide the basic and advanced features.

Hardware Requirements:

Here we list the hardware requirements for recording and processing of 365 days transmissions of 256 channels.

One hour of a 720p (i.e. a resolution 1280x720) video encoded at conservative sampling rate of 2 Mbps (2megabits per second) requires 900 MB (MegaBytes) of storage, *i.e.* 3600 x 2Mbps = 7200 megabits or 900MB. A 24 hour TV channel transmission will require

2 http://e.huawei.com/en/products/cloud-computing-dc/storage/unified-storage/2800-V3

3 http://voiceinteraction.us

4 https://www.tveyes.com/

around 24 x 900 = 22 GB of storage, whereas for storing 256 channels transmission for 24 hours will require 22 x 256 = 5632 GB = 5 TB (TeraBytes) per day. Thus per year around 365x5 =2007 TB of disk storage will be required to store the complete transmission. This is without considering the mirroring or backup of video storage.

Google Clouds[5] provide storage at $0.02 GB/month so at its peak we will be storing 2007 TB per month at the storage unit which translates to around 2007 x 0.02 x 1024 = 41103 $ per month, which is not feasible at all.

An alternative and more economical option will be to invest in off-the-shelf storage and processing solution like Huawei 2800 video processing unit. This solution costs around ~26000 $ per unit[6] and can supports large scale concurrent storage and retrieval of video data. However the cost of the attached harddisks has to be determined.

In parallel to storage requirements, there will be network bandwidth and computational processing requirements. For instance, for streaming 256 channels at 5Mbps a network connection with at least 1Gbps, and ideally 10Gbps connection will be required. Similarly, to process the incoming stream a cluster of computers with state-of-the-art GPUs and CPUs will be required to perform all the processing on the incoming stream of data. A more precise price estimate can only be established after the finalization of product features.

5 https://cloud.google.com/storage/pricing

6 http://itprice.com/huawei/2800-2c96g-8ge-ac.html

List Of Features The System Should Provide:

As augmented aired TV streams contain rich informational content thus different level of details can be extracted from the stream based on the textual, audio and visual cues. This section gives the overview of features that we are capable of building and think such a system must provide keeping in view social, governmental and commercial demands.

We categorize these features into basic and advanced groups. The basic sets of features are those that we would ideally like to built into the system. Advanced set of features are built on the top of the basic.

At the very basic level, our system should be reliably able to transcribe and tag the video frames based on the Urdu and English text present in the video frames. Although, currently no such system exists – according to our best of knowledge – however we are close to finishing such a system that can reliably transcribe a video based on the Urdu text in it.

Secondly, it should be able do transcription based on the speech using recent advancements in the deep learning. Given that now there are APIs that can be used to convert speech into Urdu text – although they are costly at the moment, *c.f.*, Google Speech API –, and the wealth of labelled data available with the TV channels it should be possible to train in-house deep learning models that can reliably transcribe the Urdu speech to text.

Thirdly, the system should be able to parse the video frames and should provide basic image based searching capabilities along with profiling and logging of different visual content being aired on the channels, *e.g.*, type of advertisements, their duration, foreign content, etc.

Here we provide a complete summary of different features that can be incorporated into system based on the audio, visual and textual information.

Summary of Features based on Visual Information:

1. Transcription of Urdu Text From Image Frames: This module will be responsible of extracting written text from the stream of input frames. This is a must have feature, and can be reliably built using recent deep learning techniques. We already have a competitive and reliable product near finishing phase.

2. Frame based retrieval: This module will allow the user to query the complete database based on a query image. Thus a user can check and track when a specific picture of a location was aired, *etc.*, Since content based image retrieval is now a mature field, this shouldn't be a problem and thus should be incorporated into the system.

3. Person based retrieval: Using this module user can query when and where a person appeared on any given channel such as in interviews, news, TV dramas, *etc*. This module will be built on the recent advancements in the face recognition technology.

4. Identification and profiling of obscene content: This module will be responsible for the detection and profiling of obscene content on the media. This can allow to detect what level of obscene content is being displayed and for how long at any given channel. All the right ingredients are there to build it and can allow to monitor the type of content being aired on specific channels.

5. Identification of foreign content: There are some channels which air both local and foreign content. This module will be responsible of doing the identification of how much of foreign content is being aired and whether it is being aired according to alloted time-limit or not.

6. Identification of commercial timings and per channel logging: This module will be responsible for identifying the number of commercials being aired and their duration on each channel.

Summary of Features based on Textual Information:

Once text has been extracted from the video (via image parsing or sound transcribing), multiple modules can be built on it:

1. Text based search: This module will enable to query the videos database based on a text query, just like a search engine. Users will be able to filter the results based on channel, day, time, *etc*.

2. Identification of person/organization from text: This module will build and deploy Urdu and English Name Entity Recognition (NER) algorithms for the detection and mentions of names, organizations, locations at any given time on any channel. This will allow to build specific triggers and services to alert any organization or person mentioned during the airing of news.

3. News sentiments & identification of hate speech: This module will be able to tell the sentiments of the aired content, whether it is positive, negative or neutral. This will make use of distributed word representations with memory-networks algorithms to decide the sentiment of news. Another subfeature of this module will be identification of hate speech whether if it contains any provocative substance inciting hate, urging violence or any radicalization content.

Summary of Features based on Audio Information:

1. Speech Matching (Finger printing of Speech): This module will enable the user to query the video corpus for extracting all the video clips that matches a specific voice pattern.

Software Requirements

All the listed capabilities can be built using the open-source tools without need of any proprietary tools whatsoever. A non-exhaustive list of languages & tools required to build the system will be as follows. Languages will include Python, Scala and C++. Libraries and tools will include Tensorflow, Keras, Opencv, Spacy, NLTK, Django, Flask, Apache Kafka, PySpark and Storm.

A Potential Processing Pipeline:

Here we propose the potential processing pipeline. The project will include individual modules for capturing, processing and analysing, and storing extracted data from video. Figure 1 gives the pictorial description of the first variant of major data engineering and processing modules.

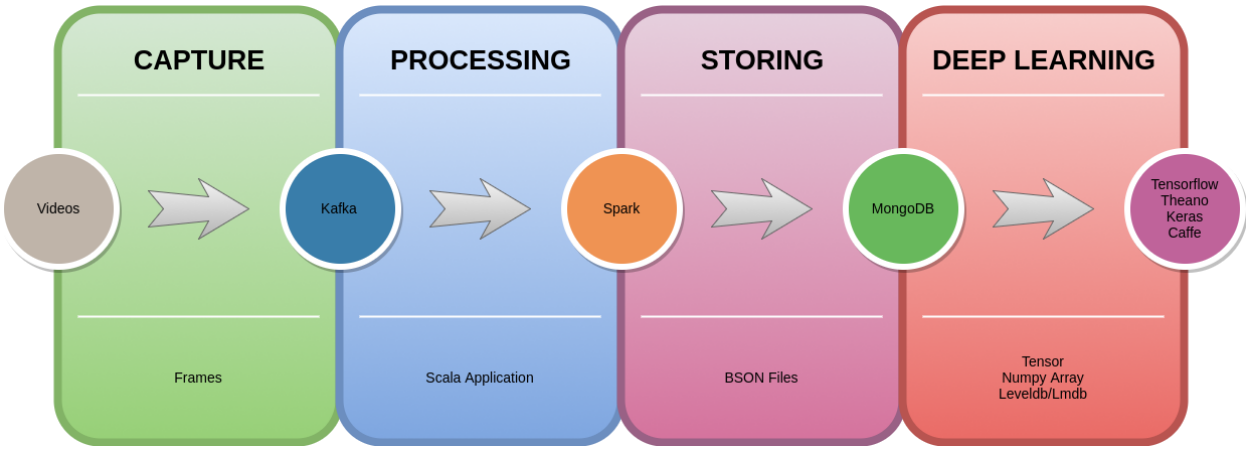

Figure 1: A Flow Graph of Stream Processing Pipeline

In capturing module, we will record and store source videos daily.

In processing and storing modules, data will be processed through Kafka and Spark with the help of data engineering and then stored into a database for future search and delivery.

In deep learning module, state-of-the-art deep learning based computer vision, natural language processing, and speech processing techniques based will be applied on initially processed data to extract more detailed level information as required according to defined features.