

Viznotes – Visual Summaries for videos

Fabin Rasheed

A Thesis Submitted to
Indian Institute of Technology Hyderabad
In Partial Fulfillment of the Requirements for
The Degree of Master of Design




भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Department of Design, IIT Hyderabad

2016

Declaration

I declare that this written submission represents my ideas in my own words, and where ideas or words of others have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources that have thus not been properly cited, or from whom proper permission has not been taken when needed.



(Signature)

Fabin Rasheed

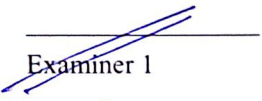
(Name)

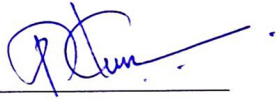
MD14MDes11001

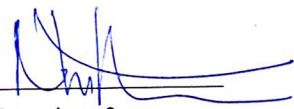
(Roll No.)

Approval Sheet

This Thesis entitled Viznotes-Visual Summaries for videos by Fabin Rasheed is approved for the degree of
Master of Design from IIT Hyderabad


Examiner 1


Examiner 2


Examiner 3

Contents

1	Introduction	5
2	Review of Literature	11
3	Study	16
4	Design	19
5	Usability Evaluation	32
6	Future Work and Conclusion	40
7	Reference	42
8	Acknowledgement	48

1 Introduction

1.1 Abstract

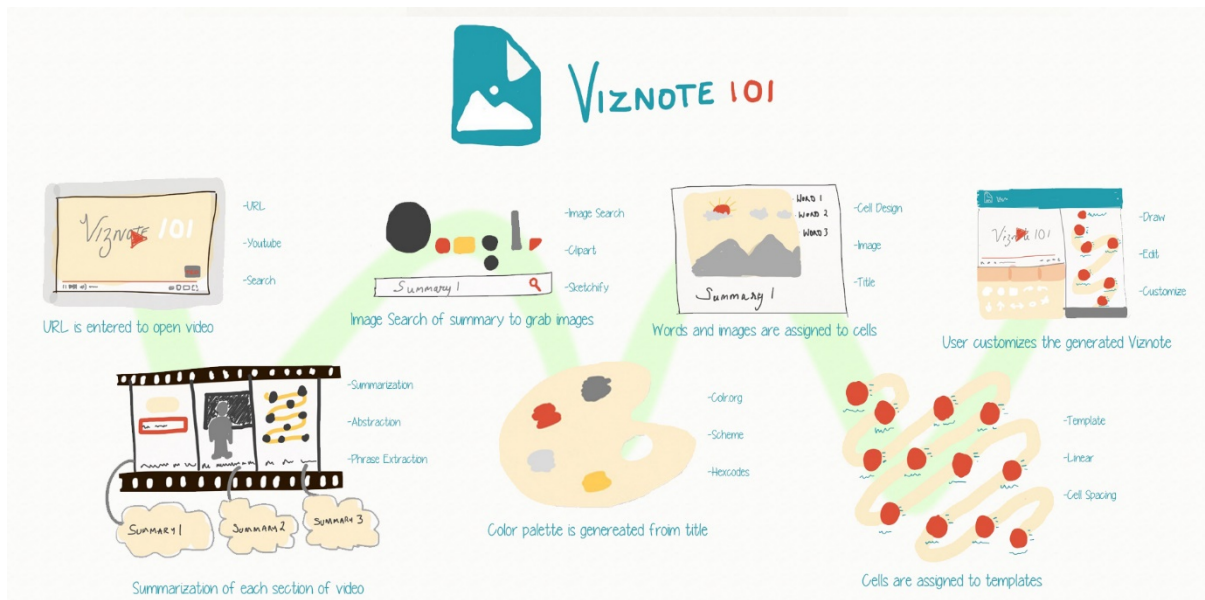


Figure 1.1 Viznote 101

This project presents a method of visually summarizing TED-like videos called Viznotes. The Viznotes interface provides a structured yet organic summarization of the contents of the video. Derived from the concepts of sketchnoting, this interface provides segments of video represented as a sketch like image with summary of the segments and keywords arranged in a pre-defined template, with certain elements showing chronology and relations. Viznote also provides an interface for navigation of the videos. Further it also enables the user to customize and make a more personal visual summary. Tools like sketching, sketch components, screen image representation etc. help users to leverage additional functionality in note taking. The project also proposes the Viznotes object model for better functionality and control over the system, and for linking it with other

multimedia systems. The notes created using Viznotes fare significantly better than unstructured sketchnotes and baseline transcript techniques in the initial content labelling experiment.

1.2 Sketching

Sketching has been traditionally used as a means of expressing ideas. Sketches are often used to tell stories or a narrative and can lead a person through a content in an active fashion. A combination of textual and visual elements could be much more appealing and informative to a person than other mediums. More benefit is received from text and sketch representations since they are seen as qualitatively different and as complementing each other [37]. In its most basic form, sketches bring out a visual explanation for a series of thoughts or concepts, which is the reason people adopt sketching or visual markers in their everyday practice.

Flemming and Mills [2] suggested the VARK model to classify types of learners, in which "visual learners" denote learners who extensively use visual representation or prefer to consume information in visual forms such as maps, diagrams, charts, graphs, flow charts, labelled diagrams, and many other visual cues. Myers & Brigg [3] place the learner personality type in the category of "INtuition". These are learners who take note of the 'big picture'; namely, how ideas connect and describe things in a more figurative or poetic fashion. Sketching as activity also provides better cognitive capabilities and is considered a befitting pedagogical tool [4]. Craft and Crains [4] present three key shreds of

evidence on the importance of sketches and sketching as an activity. They are i) the design value of sketching, ii) its importance as a cognitive support tool, and iii) its usefulness in fostering creativity. Categorically, sketches or visual notes are prepared for its quick referencing and easy recall proposition. As a medium of artistic expression and an ideation tool, sketches are being used as a tool for education in the form of visual notes, pen-based interactive tutors and concept summarizations [30].

This project presents a novel approach of summarizing videos using sketches as a medium, called Viznotes. A combination of sketches and text are arranged sequentially to form a summary of the video, which would also act as a navigation and customization interface. Sketches are generated using image processing on illustrations and the content is obtained relevant to the particular topic at hand. The project also proposes the creation of an object model for such summaries for easier manipulation and control.

1.3 Sketchnotes

Sketchnotes or visual-note-making as a summarization technique has picked up popularity, evident in the work of a group of sketchnote artists at the sketchnotearmy [7] dedicated to finding and showcasing sketchnoters around the world. Sketchnotes are being created for a wide variety of informational multimedia. Apart from lectures and conference talks, sketchnotes are also being created for article summaries, stories, historic events, and movie summarizations.



Figure 1.2 Examples of sketchnotes (© Ogilvy Notes)

While there is increasing popularity for more informational multimedia consumption, sketchnotes are only used by creative professionals, the key reasons being:

1. Sketchnote is considered a creative method of summarization, which requires the creator to be reasonably good with sketching skills and possess a versatile visual vocabulary. Hence, we find only a few sketchnote authors and publishers summarizing multimedia content.
2. Sketchnotes are perceived to be lot more fun than serious [8], largely created for a conference audience to serve as a reference and talking point. However,

people who have not attended the talk or watched the lecture video cannot take away elaborate and meaningful learnings from these notes.

3. Sketchnotes often tend to miss some important events in the course or length of the talk where content summarization critically depends on the discretion of the creator. These are more like unstructured visual summaries; hence automatic sketchnote generation is a challenging problem to solve.

However, sketchnotes find plenty of applicability in designing and creating visual summaries for multimedia content, especially TED-like informational or for that matter any tutoring/lecture videos. Prior research on the applicability of notes and annotations for learning indicate the twin processes of recording a note [16] and reviewing notes [17] as supporting learning and enhancing learner experience. Sketchnotes can be an effective medium to represent the enormous amount of Open Education Resources (OERs) present on the web. On other hand, OERs definitely require better resource indexing, presentation, management, and summarization. Extant literature indicates a good volume of work in creating multimedia summaries and browsing systems [6, 18]. Sketchnotes as a quick reference tool can further enrich multimedia consumption experiences. Additionally, an automated method of generating sketchnote-like summaries from videos could help democratize the use of such visual learning tools.

The Viznotes system creates sketchnote-like structured visual summaries of TED-like videos, and further allow learners to customize, edit the tool-generated summary from the video, allow video navigation via hyperlinks from summaries,

quick referencing and future concept revisions. For current implementation, we focus on TED videos as they cover a broad spectrum of content from technology and entertainment to design. Our interface allows the user to view key concepts presented in the video in a sketch like abstraction called sketch cells, which also consists of supporting text and key phrases. We present the Viznotes object model that allows for efficient rendering, customizing and manipulation of sketch elements through the placing of sketch cells in the user interface. The design and formatting of Viznotes leverage chronological, relational and image properties of concepts discussed in the video by a careful arrangement of sketch cells in the generated sketch template. We present the effectiveness of these Viznotes through a labelling experiment conducted with 30 users.

2 Review of Literature

Recent work has focused on improving the browsing, skimming and automatic summarization capabilities for videos of long lectures or Ted-like talks. The literature review can be classified into three key strands, which are 1). Design for note-taking and generation for computing devices, 2). Multimedia content summarization using annotations, summaries, content driven approaches, and 3). Visual narrative based summarization of videos.

2.1 Design for note-taking and generation

Note taking is a highly personal activity and each person has his or her personal style of taking notes. People generally take notes to organize their everyday plan or any acquired knowledge for future referencing. To aid note-taking, there are many commercial applications available such as Microsoft OneNote and 3M Post-it, as well tools for visual note-taking on touch enabled devices. Modern day computing device and the ubiquity of high end feature- enabled mobile devices (such as imaging and audio note) make them even more relevant note taking platforms [9]. Dynamite [29] is one such early tool that merges the benefits of paper note-taking with computers. It proposes four key properties of computer based note-taking and viewing applications. These are: i). Paper-like user interface, ii). Text keywords for content indexing, iii). Easy retrieval of specific ink & notes with dynamically changing view area, and iv). Content highlighting

capabilities. However, the extent to which note-taking interfaces interfere or influence the personal note-taking behaviour is an interesting research question. Notes also allow users to keep a track of activities in a free flow format called unstructured notes. However, it is observed that an unstructured format often leads to informational overload [10]. Thus, much of the note-taking tools we referred to are singularly targeted to producing structured notes only.

Active Notes [11] make an attempt to render unstructured notes usable by supplementing action oriented proactive interaction points. It unifies different personal information managers and integrates it to the notes for better note sharing and collaboration. Applications such as Active Notes have been designed for note-taking from scratch, and these consider user interaction as one of the key element for rendering summaries. However, in the case of automated visual note generation, it is difficult to achieve a structure while maintaining the sketch-like user experience.

2.2 Notes for multimedia content (annotations, summaries, abstractions)

Bauer, et al., [15] particularly look at the context of education and present an understanding on how people annotate documents and the implications for design of digital note taking applications. They provide evidence that copy-paste based note-taking can be more efficient than typing but can reduce attention and learning to some extent. Thus, it is important to examine the trade-offs involved in computer mediated note-taking or auto-generated notes. In case of visual notes,

user preference plays an important role. User interactions with any note-taking tool depend on the type of content that is being consumed [seen, heard or learnt]. Unlike static informational contents, videos provide the freedom to pause, replay, navigate and scrub through its length. Viewers often require a high-level overview of the topics presented in the video for quick referencing and better discovery of the most relevant content. Few of the early works such as video manga [14] and hyper video summaries [25] were developed to automatically summarize videos into pictorial representations. A similar approach was followed by Boreczky, et al. [21], which helped navigating and captioning summaries. Although this meant summarizing video pictorially, the tool relied only on screen captures and rectilinear layouts to create them.

In recent times, a lot of work focuses on developing better indexing and browsing capabilities for educational videos. Troung and Venkatesh's [6] survey of work on video summarization techniques divide methods into two main approaches; i). Using key frames to represent a sequence of important contents, and ii). Using video skim method by removing the non-important or redundant part of content. However, these summarization techniques fetch visual imagery only from the video content. In a more recent work called Video Digests [5] proposes a new format for informational videos that help users to browse and skim by segmenting videos into a chapter structure with important thumbnails and accompanying text summaries. Yadav et al. [23], and Kim et al. [24], propose a mix of content driven multi-modal interaction techniques for improving lecture video navigation.

However, both fail to provide a one-glance summary snapshot of sequence of events or concepts presented in the video. We found the tool Visual Transcript [12] to be closest to visual notes providing a method of video abstraction for chalk-and-talk style videos by extracting content from the blackboard using image-processing techniques. Visual transcript provides accompanying visual information along with the text transcript in a linear and structured manner. However, the process of rendering visual information heavily depends on visual segments written on the blackboard and tend to become very long. Thereby, they fail to serve as a quick refresher. The above-mentioned techniques serve as effective multimedia interaction systems. We believe these techniques can be improved tremendously along with the use of visual summaries.

2.3 Visual narrative based approaches of summarization

Identifying the principal, relevant and related visual content within a video is a long-standing research problem. Semantic information mining technologies and image processing have been extensively used to provide a fine-grained understanding of content in video. Studies in video classification include news video classification by Qi, et al., [19], and ontology based classification of video concepts by Wu et al. [22] Such concept detection and classification helps to generate ideas on how to classify instructional videos and select concepts from each of them. Shipman et al [27] use the property of relational hierarchies of video content to provide the most important clips as anchors in the summary. Christel

et al. [28] propose a method of automatically generating collages as dynamic summaries for news by identifying relationship through extracted metadata. This relational hierarchy and concept based similarity serves as the key to detail-on-demand based mapping of video segments. Chan et al. [26] propose a system of automatic storytelling in the game of ‘World of Warcraft’ in the form of a comic strip by analysing user logs corresponding to game screenshots. Their system aims at identifying the most important and memorable user actions in the game, which is also key to an effective visual summary generation.

3 Study

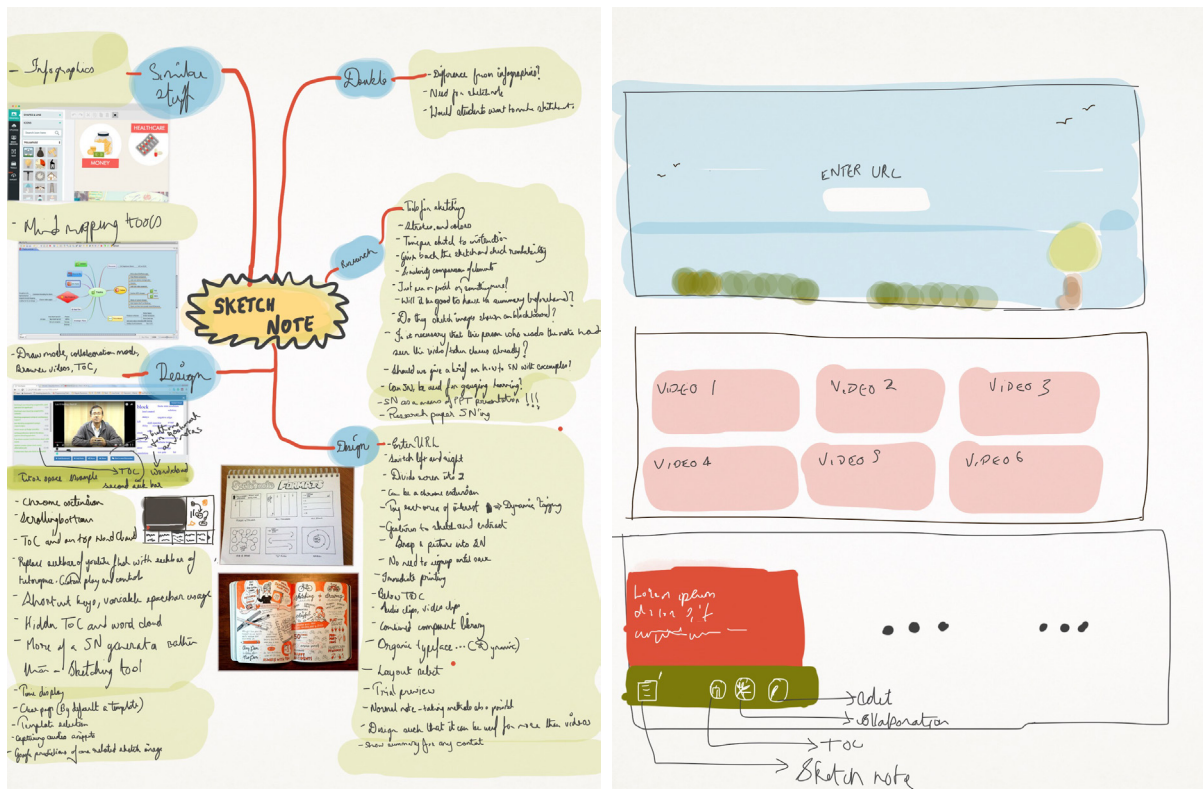


Figure 3.1 Sketchnote summary of study

From the literature review, the following design goals were identified:

- How might we generate visual notes, which are more structured, while maintaining sketchnotes-like aesthetics and user experience intact?
- How might we generate visual notes, which serve as a quick refresher and further compliment multimedia interaction techniques?
- How might we generate visual notes, while maintaining its chronological, relational properties and its salient events intact?

The initial studies started with a competitor analysis. This included reviewing popular mind-mapping tools, summarizing tools, sketching tools etc. Besides getting a grasp of how sketching by itself helps abstract ideas, these studies revealed how the design could be made better suited to different types of users. The personas of the target users were identified. (Table 3.1)

User studies on the two types of personas revealed the requirement of a better summarization and note-taking tool for Online education portals and easy summarization of other videos. It also revealed the requirement of having the same tool within the workflow of such online education sites and it should be able to provide customization. Notes taken by students are generally for the purpose of reference by the students themselves. Students should thus be able to feel their own signature in such notes.

Name	Ted Borwsky	Linda Hamilton
Age	32	27
Occupation	IT Specialist	Masters Student
Relevant Hobbies	Watching youtube videos	Regularly watches online MOOC's
Other remarks	Information hoarder, loves snippets	Has a lot of disorganised notes

Table 3.1 Personas of target users

On the other hand, people interested in viewing videos on a larger scale loves hopping through videos. In this case they require summarization of videos and sections of videos, and quick snapshots of different sections. Besides studying users who view videos, we talked to sketchnote artists and found how their regular workflow is. How much amount of summarization do they provide, how they plan the workflow before starting off, how they design the space available, what tools and techniques they use to represent ideas and what are the common sketch elements used- all these questions helped gather an idea on how the representation of visual summaries could be done and helped in designing the interface as well as the system.

The design of the system was then decided using the results of the study. This included the decision of basing the Viznote interface on a template driven design with components called sketch cells. Besides being a summarization interface, the ability to customize the Viznote using sketching tools, and sketch elements were added. The third use case included using it as a navigational interface. The system design and its flow is given in the next section.

4 Design

4.1 The system design

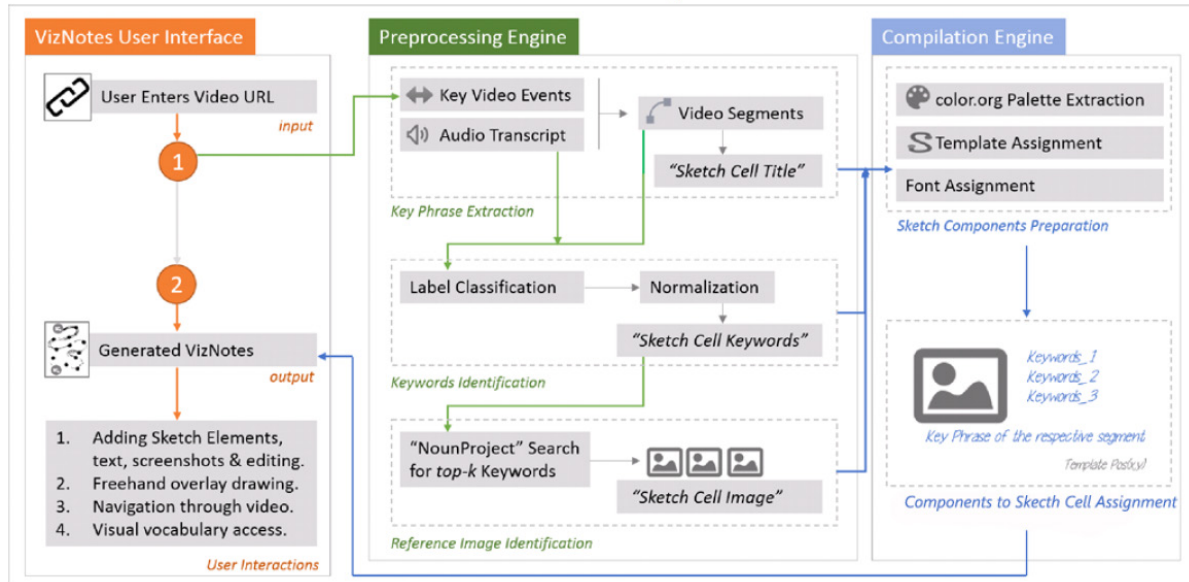


Figure 4.1 Viznotes Flow Diagram

4.1.1 Components of System

The two key components of the Viznotes system are Video Preprocessing Engine and the Viznotes compilation engine.

4.1.1.1 Video Preprocessing Engine

The Viznotes preprocessing engine passes video content from, say, a TED-like talk, through a summarizer that primarily identifies audio and visual cues inside the video. The summarizer engine provides handles for content summarization, indexing and sketch imagery identification. The key steps in Viznotes preprocessing are detailed below:

1. Key Phrase Extraction: The key phrase extraction from the video is done by an audio transcript extraction. Often times lecture and TED videos are already available with its audio transcript. These audio transcripts are present in a paragraph format along with the beginning and end timestamp. We assume these paragraphs to be key segments of the video. In case the paragraph time stamp is unknown, we identify a video segment using image processing that provide slide transitions as the key indicator. Slide transitions are assumed as the point at which the discussion in a talk changes from one context to another.

2. Keywords identification for each video segment: The extracted audio transcript for each segment of the video is then normalized. This basic text normalization essentially removes all the stop words such as ‘a’, ‘an’, ‘the’, ‘of’, and only uses the root words as keywords through Stemming. The top-3 keywords are then identified based on the frequency of occurrence.

3. Reference image identification: Once key phrase extraction from the video segment is executed, the keyword is identified and passed through a custom search for reference image identification. E.g. if a phrase like “I specialize in human behavioral research, and applying what we learn to think about the future in different ways, and to design for that future” is obtained, the corresponding search term is identified as “behavior” or “future”. For each keyword, up to 5 top results of images are retrieved.

The main reason for using keywords instead of the key phrases is to avoid redundancy and irrelevancy of image search results. We first used the Google

custom search API for clipart image search. However, search results for some of the keywords were very broad and did not specifically represent the context underlying the video segment. To overcome this broad nature of search, we appended the search term with keywords identified from the subsequent video segment. For example, if the search term for first video segments is “behavior” and search term for the next is “people”, we considered the search term for the first as “behavior + people”. The initial aim was to obtain a pairwise approximation of related video segments, and thereby obtaining better search results. However, the above-mentioned technique resulted in all of the video segments having similar reference images across varied key word search terms. Thus, currently we are using the existing NounProject [1] library, which has icon images present with relevant tags. Using the tag based search on NounProject, we retrieve top-5 reference images for every video segment. This, we found, effectively represents the context in a video segment.

Post retrieval, the images are formatted into a sketchy form. This requires the images to be processed into two layers. The first layer helps threshold the image and provides two sets of colors to the image. Then a pattern is overlaid on the major color. The second layer is then processed to obtain image, to which a darker color is assigned. The edge-image is overlaid on the top of the previous image. Both the layers are combined to obtain the final sketchy reference image.

4.1.1.2 Viznotes Compilation Engine

The Viznotes compilation engine accesses all sketch elements retrieved during runtime, and fetches it for Viznotes rendering. Viznotes rendering requires sketch objects to be prepared for each video segments. We call these video segment-specific sketch representations as Sketch Cells (SC). These SCs are rendered based on a predefined object model along with the key entities of a reference image, the title phrase and the supporting keyword labels. The number of SCs is obtained from the number of video segments.

Similar to the document object model, we provide a logical structure to Viznotes, which can contain more than one SCs. Once SC objects are identified, the SC anchor points are computed such that it can be overlaid on pre-defined layouts in SVG format. In the current implementation, we have designed three of such layouts namely fluidic, organic and linear as shown in Figure 4.2. We call these layouts as Sketch templates (ST). The SC, being the key object in the model, can follow various types of dynamically assigned STs. In the case of longer videos, the length of the template is dynamically increased depending on the number of SCs. The coordinates of the SCs are then designed based on the ST path length such that they are equidistant and have a breathing space between them. In future, we aim to identify the most appropriate ST based on video properties such as the context of the topic, speaker movements, emotional classification of visual cues or the audio transcript.

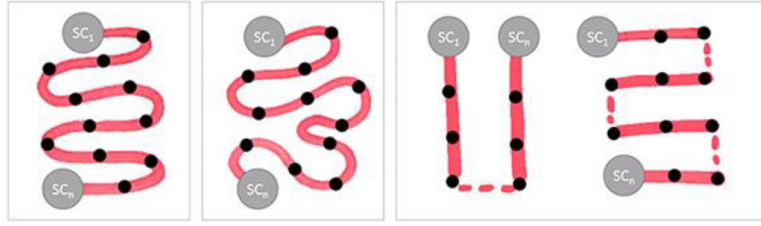


Figure 4.2 Sketch Templates: Fluidic, Organic, and Linear (from left to right)

4.1.2 Object Model

The object model encompasses not only the structure of the Viznote, but also its relational and chronological attributes. The different relationships in a video segment could be presented in Viznotes by manifesting the object attributes. The Viznote object model as shown in Figure 4.3, allows easier document manipulation when a user interacts with Viznotes. Some of the key feature achieved with this object model are:

1. The adjacency of the SCs along the ST presents the chronological relationship of the video segments.
2. If SC_i and SC_{i+n} are related, they could be shown along with a connector object.
3. If SC_i and SC_{i+m} present the same context, they could be shown with similar color scheme or highlighting style.
4. The different attributes and elements can be accessed in runtime and can be modified by “id” based referencing. Eg: If a SC has an id= “cellOne”, then sub-elements like image (img) and their attributes (like size) can be changed by

referring to the id cellOne (like cellOne.img = “xyz.jpg” or cellOne.img.size = “120,120”).

5. Each major sketch element can be customized taking care of corresponding automatic changes in sub-elements. For example, if a ST is changed from fluidic to organic, the SC positions are also changed. Positioning is responsive and change according to the screen resolution. This is achieved by not allowing any overlay of SCs with one another.

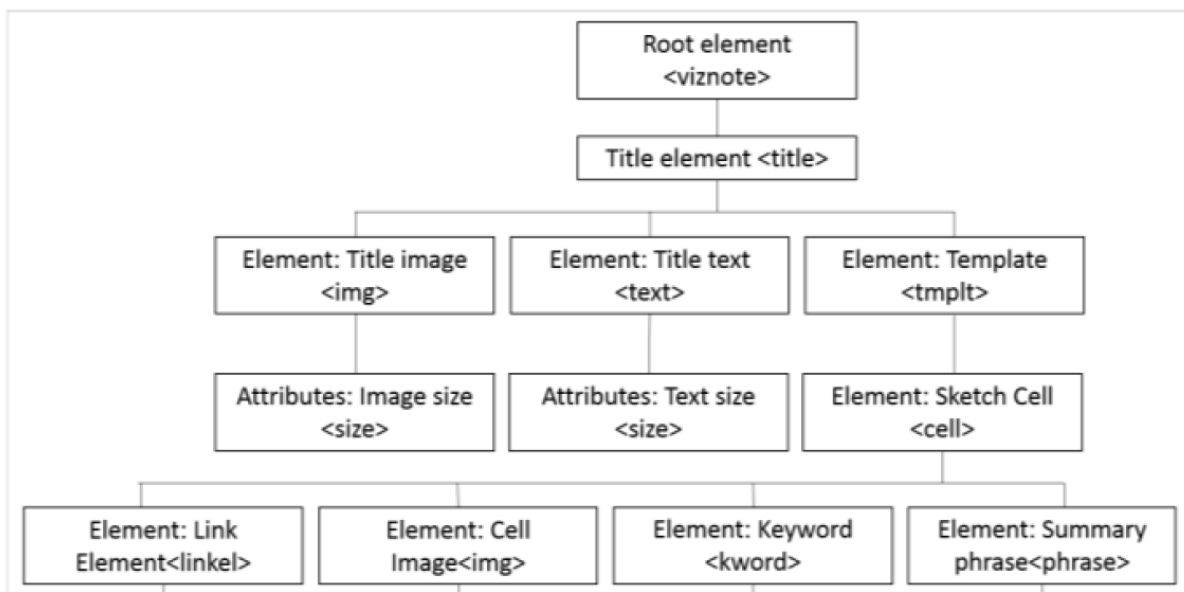


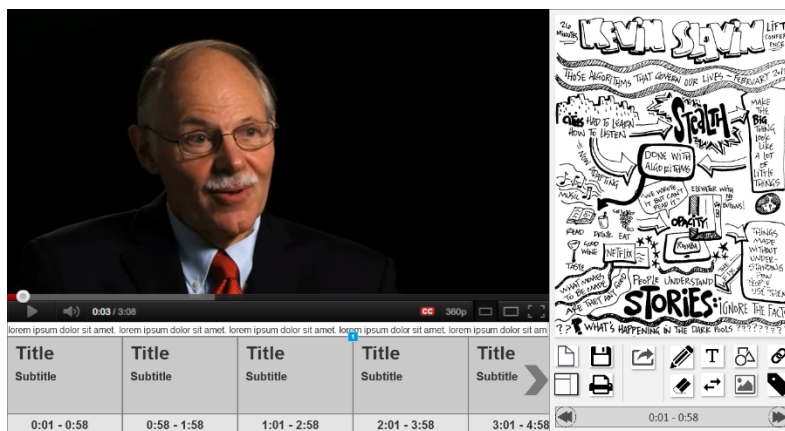
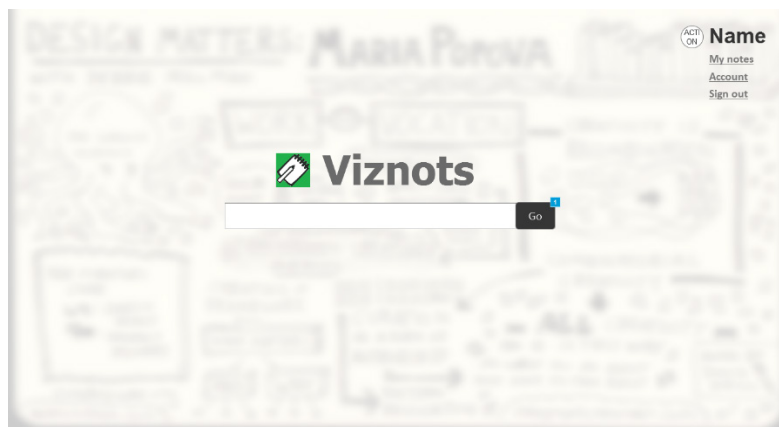
Figure 4.3 Viznotes Object Model Diagram

Colors are also dynamically obtained using keyword based color search with context. For example, a tag search for keyword “Sky” would return the hexcode “#8abceb”. The obtained colors are separated in terms of brightness and the darker color assigned to texts and edges while the lighter color is used as a fill color of the reference images.

4.2 Interface Design

4.2.1 Concepts

The initial concept included a quick reference sketchbook mounted to the side of the video being played. The video navigation was segmented and separately shown below the video. These were later combined together. The initial concepts also had multiple viznotes for a single video rather than a continuous flow version and had tools available inline taking up real estate. In the final UI, a continuous flow model was adopted and a FAB button was introduced for additional tools.



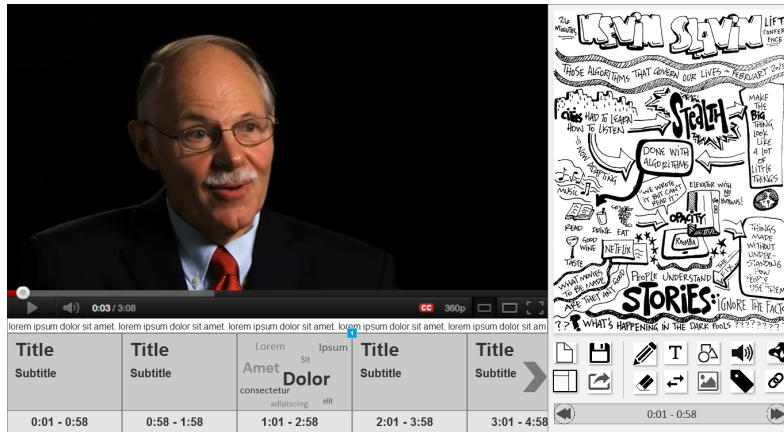
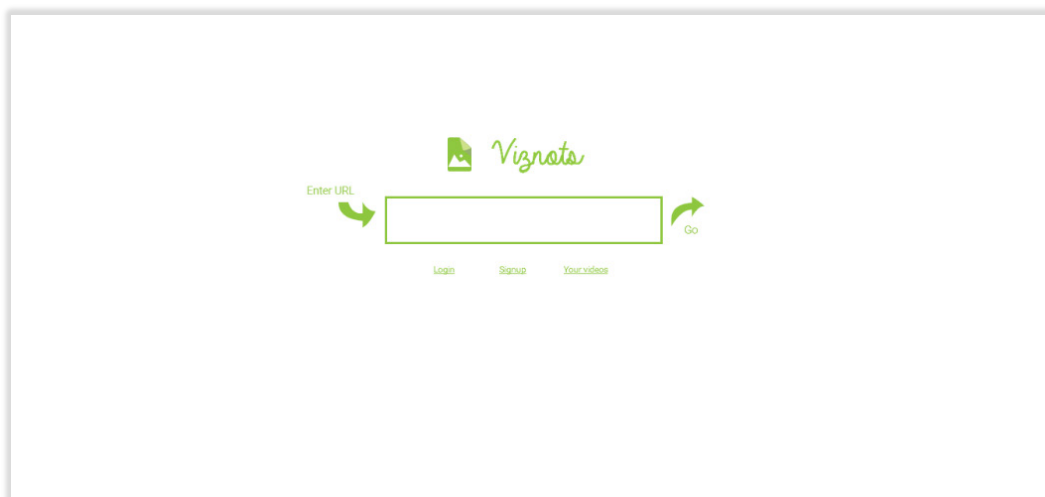


Figure 4.4 Initial wireframes

4.2.2 Final interface design

The initial screen prompts users to input a video URL. This is followed by preprocessing and Viznote compilation. The Viznote screen now displays the video and sketch area as shown in Figure 4.5. Our current implementation only accepts a TED talk with its available YouTube URL. The screen is divided into four sections: Video Player, Sketch Components, and Viznotes Viewer.



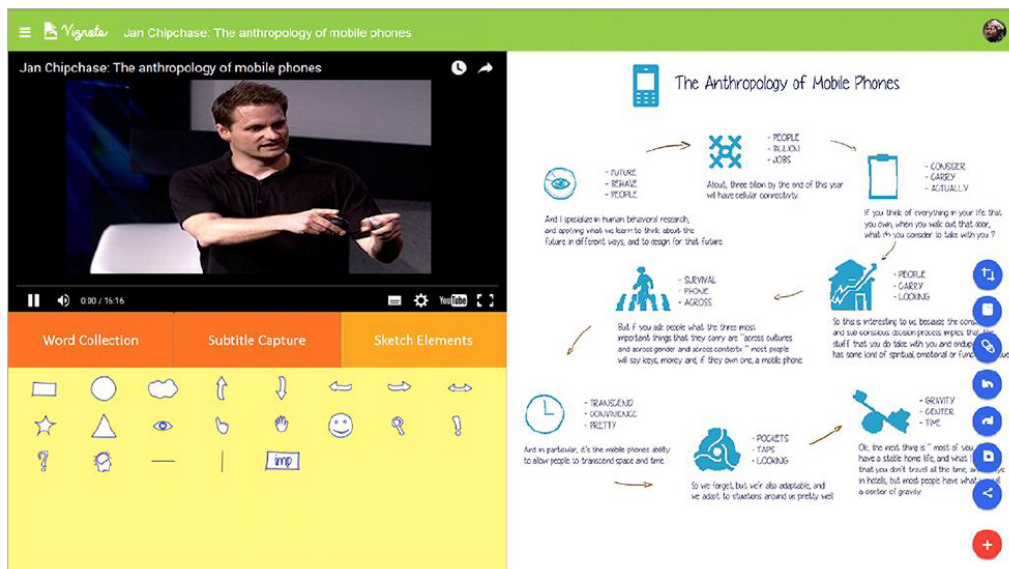


Figure 4.5 Viznote user interface

The initial screen of the Viznotes system asks for a URL to be entered, which will process the video on submission. “My viznotes” is a screen where you can see all the Viznotes a user has edited and saved. The focus of this project was primarily on the third screen which is the main Viznote editing and using interface. This is divided into 3: Video Player, Sketch Component, Viznotes viewer/editor.

4.2.2.1 Video Player

The video player consists of the playable video, where a user could pause, play and scrub at any time. For the purpose of prototyping, we selected videos with available YouTube link. Each SC carries its respective video segment time stamp as a seeking point. The video player area also serves for the capturing of a specific object, which could be a formula, diagram or any other pre-defined element in the video using screen capture.

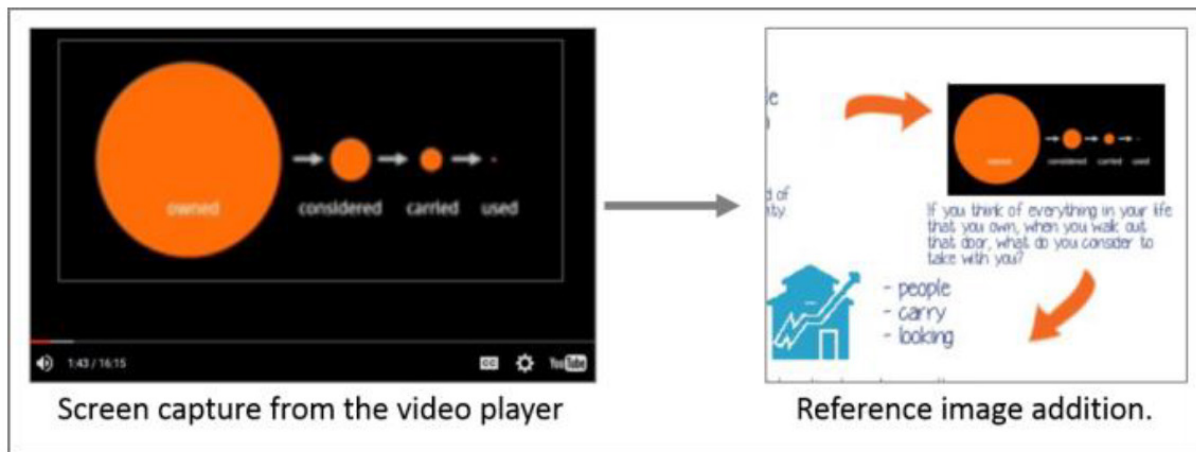


Figure 4.6 Reference image addition to the notes

4.2.2.2 Sketch Components

The sketch components section is divided into 'word collection', 'captured subtitles', and a 'sketch element component' library. The top-5 keywords identified from each video segments are displayed as word collection, and the transcripts are displayed as captured subtitles. Both keywords and sentence-wise

transcripts can be dragged and dropped into the Viznotes viewer area. The sketch elements are displayed as small icons under sketch elements section.

For the design of sketch elements, we looked at sketching tools such as Piktochart or Visio, which has a predefined classification of graphics such as lists, cycles, process, shapes, lines and many more. However, in the current implementation we have employed the Microsoft Smart Art [20] provided by Microsoft Office, for its familiarity and widespread acceptance. We also referred to sketchnotes library by OgilvyNotes [36] to identify the commonly used sketch elements.



Figure 4.7 Sketch Components

4.2.2.3 Viznotes viewer/editor

The sketch area consists of a canvas onto which different layers are rendered such as the compiled Viznote, pencil layer, erase layer, sketch element layer, screen

capture layer, etc. The screen has been designed to be responsive for screen size compatibility.

Each sketch cell corresponds to a different segment of the video. Clicking on the sketch cell will play the video corresponding to that particular sketch cell's timing. Thus, it acts as a navigation interface. We provide stylus and mouse based interaction capabilities, for Viznote-taking. This allows users to drag and drop any sketch component or screen capture layer in the Viznotes. Additionally, users can also add a link to any of these sketch components for quick video navigation for a later point in time. On dragging and dropping some sketch elements (like box, circle or a cloud), an optional text box appears enabling users to enter text into the element. A sketch-like font called "yummy cupcakes [34]" is used for rendering the text. The default images displayed on the SCs can be changed by clicking and holding to reveal an optional set of 5 top image search results of the same keyword. This enables the user to replace these SC images based on preference. We also prototyped a feature which suggests sketch components while the user is drawing using the Viznotes tool. The user will have the option to draw basic outline strokes and press a key. This user stroke will be processed in a neural network and similar images will be shown as options for the user to select from. For example, a user drawing the outline of a bird will be presented with alternate images within the class to select from. This feature would effectively augment the user's visual vocabulary nudging them to draw richer

visual notes. Further customization capabilities are also included like erase everything, move and re-position sketch cells etc.

User controls are presented on clicking the floating action button (FAB) present at the right bottom of the sketch viewer. In our early observations, we identified users preferring to pause the video even while making notes on paper. Hence, we implemented the video pausing feature on click of the FAB. The user controls have anchors for screen capture, clear screen, link, undo, redo, save (to “Your Videos”) and share Viznotes (as PDF, mail).

A material design theme was adopted for the visual design. This was such that it could be integrated with the android ecosystem especially targeting tablet interfaces where, the Viznote interface would find its best usages.

5 Usability Evaluation

5.1 Study Design

We conducted an initial labeling experiment with 30 participants to identify the effectiveness of the Viznote summaries. The participants were recruited using a convenience sampling. Participants included 20 males and 10 females, in the ages between 22-35. The participants were all users of informational multimedia content in some form such as TED talks, MOOCs, or other tutorial videos. Nine of the participants were professional UI/UX designers, who understood sketchnotes but had no experience with sketchnotes-like visual summaries for any information consumption. The aim of the labeling task was to examine the effectiveness of the quick summary provided by Viznotes in comparison with unstructured sketchnotes and baseline video transcripts. A keyword-labeling task was chosen to identify how users abstract a group of concepts present in any informational summary. Three TED videos, each five to ten minutes long were selected for the experiment. The selected videos are of high resolution and transcript enabled. Most importantly, none of the videos were ever viewed by the 30 participants. The unstructured sketchnotes for the selected TED videos were obtained from TED talk sketchnotes [13]. The baseline label list for each video was prepared with the help of another set of 10 participants. These 10 participants were asked to watch each video and provide 10 keywords discussed in the content with a weightage value on a coarser scale of 1 to 5, where 5 represents the most

important keyword. Based on the baseline labels obtained from these ten participants and considering the weightages, we obtained a ranked list of 25 labels per video. The ranked list, qualitatively and organizationally, represented the video content in a precise manner.

For each video (Vid_n), a set of three summaries were prepared; one from the summary created by our Viznotes (V_n), second from Sketchnotes (S_n) obtained from TED sketchnotes, and text transcripts embedded in the video(T_n). The 30 participants for labeling activity were divided into 3 equal sets of 10 participants each. Each set is provided with the three different type of summaries accompanied by 25 baseline labels. E.g. Set 1 was provided with 3 different images containing V₂, S₁, and T₃.

The occurrence of the 25 labels in the survey questionnaire was randomized to avoid any bias introduced by the default baseline ranking. No video was shown to the participants during the experiment and each image was shown for only a limited duration of 2 minutes. Participants were asked to select 10 keyword labels from the available list of 25 labels. The control parameter of 2 minutes was maintained to observe the effectiveness of quick referencing/recall capability. Qualitative feedback was recorded at the end of each session. We did not use any content specific Q and A approach as feedback on the videos seen by our participants as both sketchnotes and Viznotes are summarizing tools and do not intend to present the detailed concept or story presented in videos.

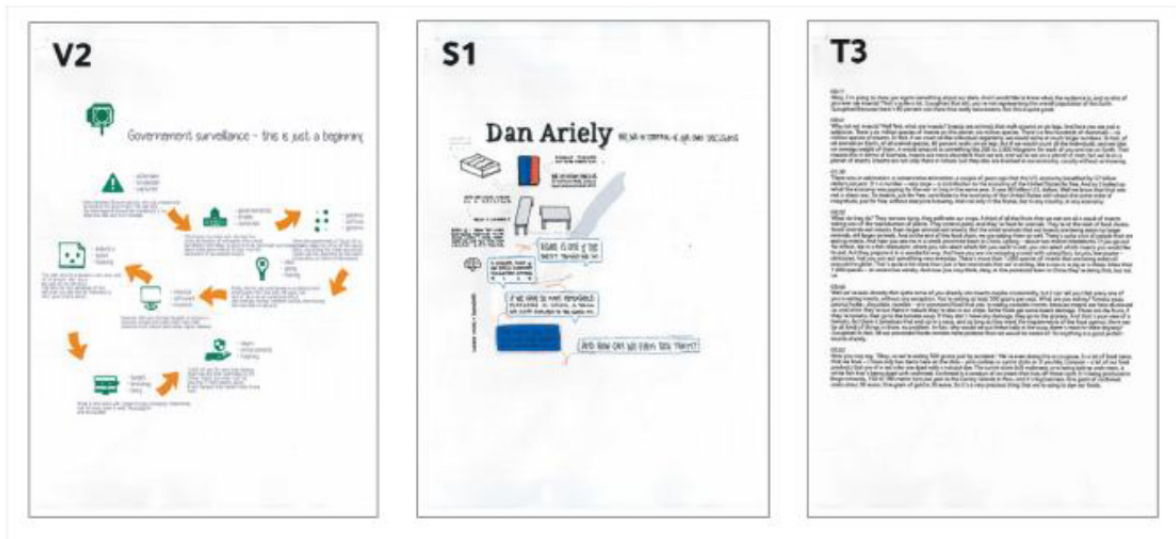


Figure 5.1 Test Set: Viznote, Sketchnote and Transcript

Before we move on to describing the methods adopted in our experiment, here are a few broad yet key questions informing our data collection:

1. Do Viznotes with its visual summary and key phrases match the human quality of summarizing and internalizing video information?
2. What is the effectiveness of abstracted visual summaries with respect to content knowledge obtained from a full video preview?

5.2 Method of observation

The baseline labels obtained from the video preview was used to observe the difference in number of label occurrences with respect to different type of summaries. We compare the total number of label occurrence from Viznotes, sketchnotes, and transcripts by using a label density plot. During the task, we also noted if users were able to complete reading the textual notes in the stipulated

two minutes and if they needed to refer back to the note while selecting their 10 best labels.

5.3 Findings

All the participants found visual summaries (both Viznotes and Sketchnotes) to be refreshing and quick, but Viznotes were preferred for presenting a structured overview. Interestingly, none of participants were able to finish reading the transcript in the given time limit of 2 minutes. In this section, we report our findings about the effectiveness of Viznotes, and our understanding of user preferences obtained from the qualitative feedback session.

5.4 Effectiveness of Viznotes

Figure 5.2 presents the label density graph obtained from sketchnotes, Viznotes, and transcripts with respect to the baseline list of labels. We observe that the top-20% of labels were obtained 99 times in case of Viznotes in comparison to 86 times in case of transcripts, followed by 85 times in case of sketchnotes. Viznotes fare better for the following top-20% of the labels as well with a total of 78 instances.

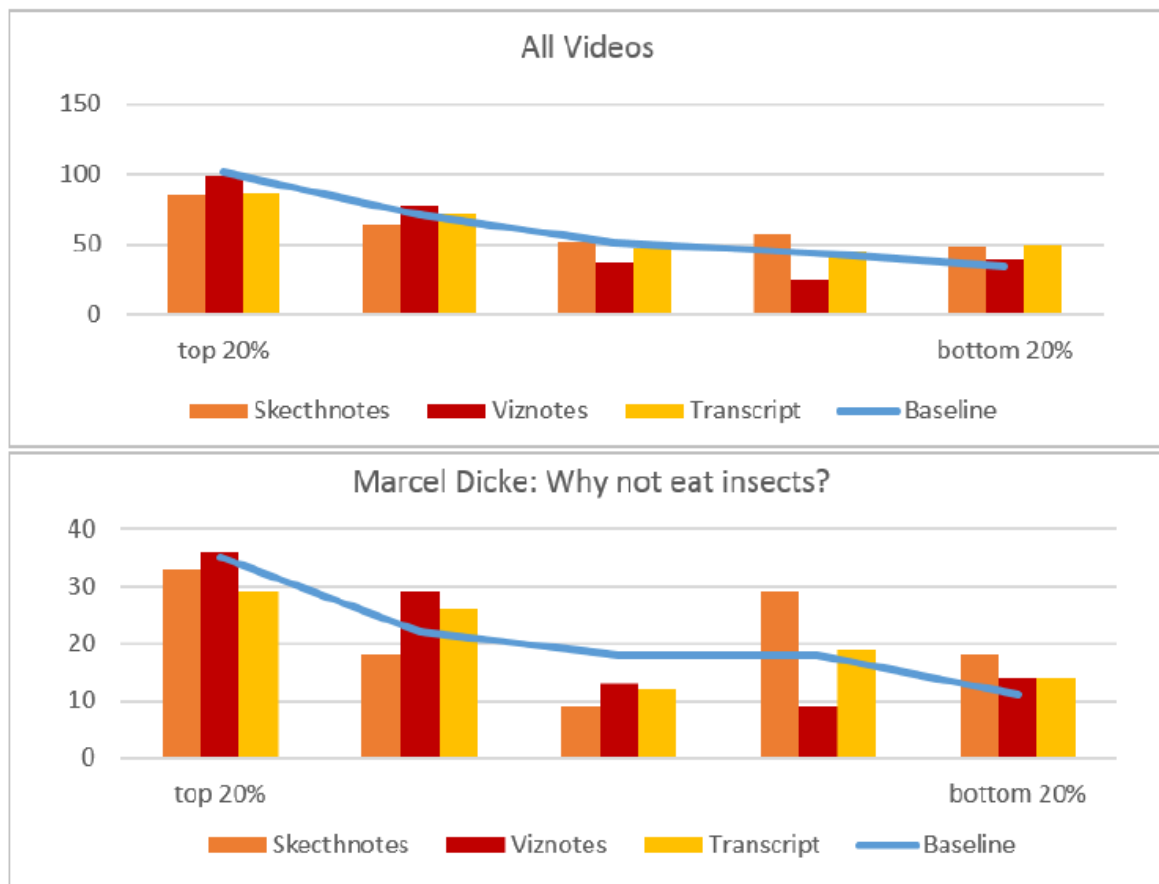


Figure 5.2 Label Occurrence Density

A higher number of occurrence essentially informs that participants were much easily able to register the top labels corresponding to the content in case of Viznotes. The bottom-20% labels represent detail specific labels such as “pollinate” and “cochineal”, which were obtained from the video “Marcel Dicke: Why not eat insects?”. A lower number of occurrence in case of Viznotes shows that participants could not find such detail- specific labels. Perhaps, in many cases, it is difficult to automate the identification of visual representation for such specific content. However, we found, with user mediation this can be achieved. Sketchnotes surprisingly provided similar or better results in comparison to the

transcripts for the bottom-20% labels, which show that manually prepared visual notes were able to present specific label details in a clearly highlighted manner. As far as quick referencing is concerned, the results validate Viznotes as effectively presenting the meta-level overview in a very short time. We also looked at the frequency of occurrence for the top-10 labels for respective videos. These top-10 labels were obtained from baseline ranking through video previews.

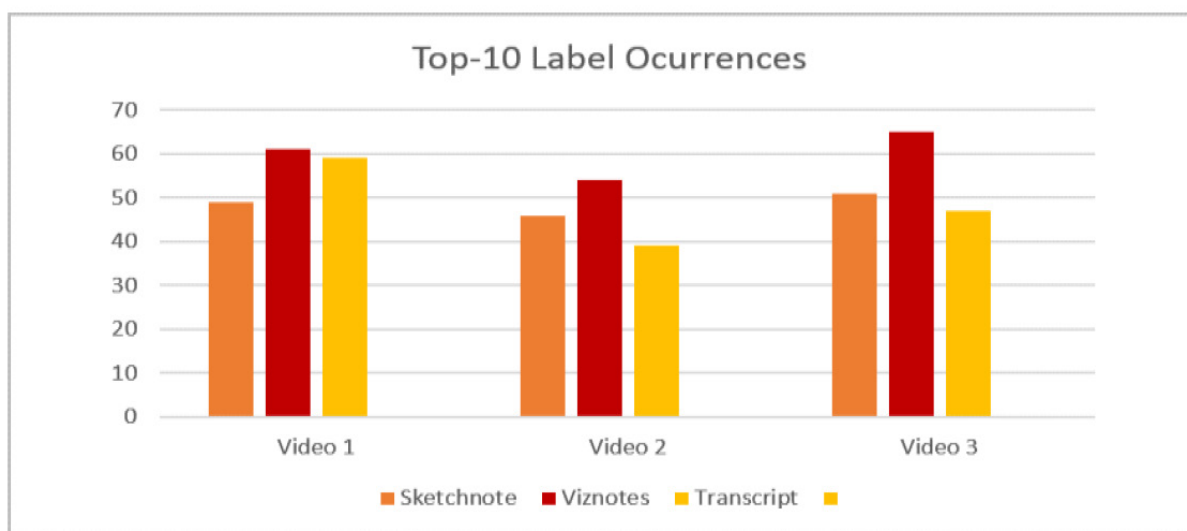


Figure 5.3 Top-10 Label Occurrences across videos

Our initial assumption remained that an effective visual abstract would present all the top-10 labels across users. Figure 5.3 shows that all the visual summaries fare better than transcripts except in the case of Vid1: “Dan Ariely asks, Are we in control of our decisions?”. It is important to note that the video is about behavioral economics, where the speaker touches upon a variety of disparate topics in the talk, but forging a cohesive story. In our qualitative feedback session, many users informed they could not understand the entire context of V1.

Interestingly, the participant who received the transcripts for Vid1 didn't complain of the same. This also indicates future challenges of abstracting certain type of contents which are more unstructured in nature and having an unclear relationship between video segments, factual accounts, and non-linear story. In future, we aim to study a corpus of such videos and derive a set of fundamental classificatory on the basis of video properties such visual imagery, storytelling, narrative and instruction style for appropriateness of visual summarization.

5.5 Attributes of effective visual summaries

Unlike sketchnotes, Viznotes were perceived to be a more formal way of visual summarization. Participants indicated that Viznotes provide a very good overview of video content but at times fails to present the whole story or factual components discussed in the video. While going through the Viznotes, participants expressed their interest in watching the full video post the labeling activity, this particularly happened in the case of Vid1 as mentioned earlier. Importantly, the use of full sentences as key phrase created some confusion among participants, as they were not able to spell out the flow of the video from one sketch cell to the next. However, it generated curiosity among our participants to watch the video later. We also received important feedback on the visual design aspects of the Viznotes. Many participants indicated towards the need of a more legible font. One of the users mentioned, "Sketchy doesn't mean it should be unreadable..., it is a tradeoff, though". In another case, a UX designer

while labeling a Viznote with one of the SC images of “Skype logo” could not notice the image. Later she exclaimed, “for me the blue color of skype logo is more important”. Often images that are also logos can render a definite meaning with its actual colors intact. Employing or customizing a color from a palate color scheme could cause some loss of information. Another example could be a stop-signal clipart image rendered on grayscale resulting in meaningless information. Three of the participants mentioned that visual summaries could replace email newsletters which, often get overlooked.

6 Future work and conclusion

In this project, we presented Viznotes that effectively summarizes TED-like informational video content and provides a quick summary viewing experience.

The object model used for Viznotes provides easy customization capabilities and effective sketch manipulation based on user interactions. The initial labeling experiment done with auto-generated Viznotes shows that it fares better significantly in comparison to textual transcripts and unstructured sketchnotes. Our exploration of emulating human intelligence for creating visual summaries opens up a whole range of research questions and future possibilities.

Being an effective summarization tool, Viznotes have the potential to be applied for other mediums as well such as documents, eBooks, and varied forms of video.

A central repository of Viznote collections would help democratize the use of visual notes for multimedia content consumption. Viznotes can be used as an alternative form of abstracts of research papers, with navigational links embedded in them hyperlinking relevant portions of the paper. Tutorial videos and books could also benefit from such abstraction and navigational links. Better summarization can be done by providing narrative and visual analysis. Online news content could also be summarized into Viznotes or Viznote-snippets. Continuous learning of usage patterns and note-taking behavior of users on Viznote will help develop better summaries and visuals. Considering user interactions and learning from usage patterns, standard Viznotes design patterns

could be developed, thereby bringing in collaborative learning into the application. Document Object Model based actions like navigation, design and search would help in increasing content accessibility. Viznote also has the possibility of being implemented as a summarization of webpages. The summaries could be displayed along with search results to give a one shot glance of the whole page. Learning large sets of images and grouping them by labels will help in better reference image suggestions for sketch.

As next step, we aim to do a detailed evaluation of the interface itself and understand the different Viznotes customization and usage patterns. An interesting direction would be to extend our studies with Viznote summarizing for different types of videos and other mediums such as documents and webpages. We foresee, visual summarization techniques being applied to various mediums conveying and indexing large sets of informational multimedia.

7 References

1. Noun Project. Noun Project, 2016. <https://thenounproject.com/search/>.
2. Fleming ND. I'm different; not dumb. Modes of presentation (VARK) in the tertiary classroom. In: Zelmer A, editor. Research and Development in Higher Education, Proceedings of the 1995 Annual Conference of the Higher Education and Research Development Society of Australasia (HERDSA). Vol. 18. Higher Education Research and Development; 1995. p. 308-13.
3. Myers, IB. & McCaulley MH. (1985) Manual, A Guide to the Development and Use of the Myers-Briggs Type Indicator, Consulting Psychologists Press, California
4. Craft, B. & Cairns, P. Sketching Sketching: Outlines of a Collaborative Design Method. , In Proc of HCI 2009 – People and Computers XXIII, pp. 65-72.
5. Amy Pavel, Colorado Reed, Björn Hartmann, and Maneesh Agrawala. 2014. Video digests: a browsable, skimmable format for informational lecture videos. In Proceedings of the 27th annual ACM symposium on User interface software and technology (UIST '14). ACM, New York, NY, USA, 573-582.
6. Ba Tu Truong and Svetha Venkatesh. 2007. Video abstraction: A systematic review and classification. ACM Trans. Multimedia Comput. Commun. Appl. 3, 1, Article 3 (February 2007).
7. Rohde, M. "About Sketchnotes - A Showcase Of Sketchnotes". Sketchnotearmy.com. N.p., 2016. Web. 20 Mar. 2016.

8. Irgens, E. and Elisabeth, I. "How To Get Started With Sketchnotes – Smashing Magazine". Smashing Magazine. N.p., 2014. Web. 20 Mar. 2016.
9. Shwetak N. Patel, Julie A. Kientz, Gillian R. Hayes, Sooraj Bhat, and Gregory D. Abowd. 2006. Farther than you may think: an empirical investigation of the proximity of users to their mobile phones. In Proceedings of the 8th international conference on Ubiquitous Computing (UbiComp'06), Paul Dourish and Adrian Friday (Eds.). Springer-Verlag, Berlin, Heidelberg, 123-140.
10. Starr R. Hiltz and Murray Turoff. 1985. Structuring computer-mediated communication systems to avoid information overload. *Commun. ACM* 28, 7 (July 1985), 680-689.
11. Suresh Chande, Panu Vartiainen, and Kimmo Rämö. 2007. Active notes: context-sensitive notes for mobile devices. In Proceedings of the 4th international conference on mobile technology, applications, and systems and the 1st international symposium on Computer human interaction in mobile technology (Mobility '07). ACM, New York, NY, USA, 716-723.
12. Hijung Valentina Shin, Floraine Berthouzoz, Wilmot Li, and Frédo Durand. 2015. Visual transcripts: lecture notes from blackboard-style lecture videos. *ACM Trans. Graph.* 34, 6, Article 240 (October 2015), 10 pages.
13. Tedsketchnotes.tumblr.com, 2016. <http://tedsketchnotes.tumblr.com/>.
14. Uchihashi, Shingo, et al. "Video manga: generating semantically meaningful video summaries." Proceedings of the seventh ACM international conference on Multimedia (Part 1). ACM, 1999.

15. Aaron Bauer and Kenneth R. Koedinger. 2008. Note-taking, selecting, and choice: designing interfaces that encourage smaller selections. In Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (JCDL '08). ACM, New York, NY, USA, 397-406.
16. Marshall, C. Towards an Ecology of Hypertext Navigation. Hypertext 1998, ACM Press (1998), Pittsburgh, PA. 40-49.
17. Marshall, C.C. & Bernheim Brush, A.J. Exploring the Relationship between Personal and Public Annotations. Proc DL 2004, ACM Press (2004), 349-357.
18. Brezeale, Darin, and Diane J. Cook. "Automatic video classification: A survey of the literature." Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 38.3 (2008): 416-430.
19. Qi, Wei, et al. "Integrating visual, audio and text analysis for news video." Image Processing, 2000. Proceedings. 2000 International Conference on. Vol. 3. IEEE, 2000.
20. "Create A Smartart Graphic - Office Support". Support.office.com. N.p., 2016. Web. 22 Mar. 2016.
21. Boreczky, John, et al. "An interactive comic book presentation for exploring video." Proceedings of the SIGCHI conference on Human Factors in Computing Systems. ACM, 2000.

22. Wu, Yi, Belle L. Tseng, and John R. Smith. "Ontology-based multi-classification learning for video concept detection." *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*. Vol. 2. IEEE, 2004.
23. Kuldeep Yadav, Kundan Shrivastava, S. Mohana Prasad, Harish Arsikere, Sonal Patil, Ranjeet Kumar, and Om Deshmukh. 2015. Content-driven Multimodal Techniques for Non-linear Video Navigation. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, NY, USA, 333-344.
24. Juho Kim, Philip J. Guo, Carrie J. Cai, Shang-Wen (Daniel) Li, Krzysztof Z. Gajos, and Robert C. Miller. 2014. Data-driven interaction techniques for improving navigation of educational videos. In *Proceedings of the 27th annual ACM symposium on User interface software and technology (UIST '14)*. ACM, New York, NY, USA, 563-572.
25. Girgensohn, Andreas, Frank Shipman, and Lynn D. Wilcox. "Hypervideo summaries." *ITCom 2003. International Society for Optics and Photonics, 2003*.
26. Chia-Jung Chan, Ruck Thawonmas, and Kuan-Ta Chen. 2009. Automatic storytelling in comics: a case study on World of Warcraft. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems (CHI EA '09)*. ACM, New York, NY, USA, 3589-3594.
27. Frank Shipman, Andreas Girgensohn, and Lynn Wilcox. 2003. Generation of interactive multi-level video summaries. In *Proceedings of the eleventh ACM*

international conference on Multimedia (MULTIMEDIA '03). ACM, New York, NY, USA, 392-401.

28. Michael G. Christel, Alexander G. Hauptmann, Howard D. Wactlar, and Tobun D. Ng. 2002. Collages as dynamic summaries for news video. In Proceedings of the tenth ACM international conference on Multimedia (MULTIMEDIA '02). ACM, New York, NY, USA, 561-569.

29. Schilit, Bill N., Lynn D. Wilcox, and Nitin Nick Sawhney. "Merging the benefits of paper notebooks with the power of computers in dynamite." CHI'97 Extended Abstracts on Human Factors in Computing Systems. ACM, 1997.

30. Joseph J. LaViola. 2007. Sketching and education. In ACM SIGGRAPH 2007 courses (SIGGRAPH '07). ACM, New York, NY, USA, , Article 6 .

31. Lothian, N. Classifier4J - Classifier4J. Classifier4j.sourceforge.net, 2016.

32. Kuldeep Yadav, Kundan Shrivastava, and Om Deshmukh. 2014. Towards Supporting Non-linear Navigation in Educational Videos. In Proceedings of the 16th International Conference on Multimodal Interaction (ICMI '14). ACM, New York, NY, USA, 82-83.

33. colr.org API. Colr.org, 2016. <http://www.colr.org/api.html>.

34. Yummy Cupcakes Font - 1001 Free Fonts. 1001 Free Fonts, 2016 by <http://bythebutterfly.com> http://www.1001freefonts.com/yummy_cupcakes.font.

35. Schlegel, A. processing GUI, controlP5. Sojamo.de, 2016. <http://www.sojamo.de/libraries/controlP5/>

36. Ogilvy Notes. Ogilvynotes.tumblr.com, 2012. <http://ogilvynotes.tumblr.com/>.

37. Claudia Leopold, Detlev Leutner. Science text comprehension: Drawing, main idea selection, and summarizing as learning strategies. *Learning and Instruction*, Volume 22, Issue 1.

8 Acknowledgements

I thank Dr. Prasad Onkar for his valuable guidance in this project. I also thank Xerox Research Centre India, for giving me the opportunity to do this project. I would also like to express my heartfelt thanks to my mentor Mr. Jyotirmaya Mahapatra, and to Mrs. Nimmi Rangaswamy and Mr. Saurabh Srivastava for their pointers and guidance.