Automatic Generation of Two-Level Hierarchical Tutorials from Instructional Makeup Videos

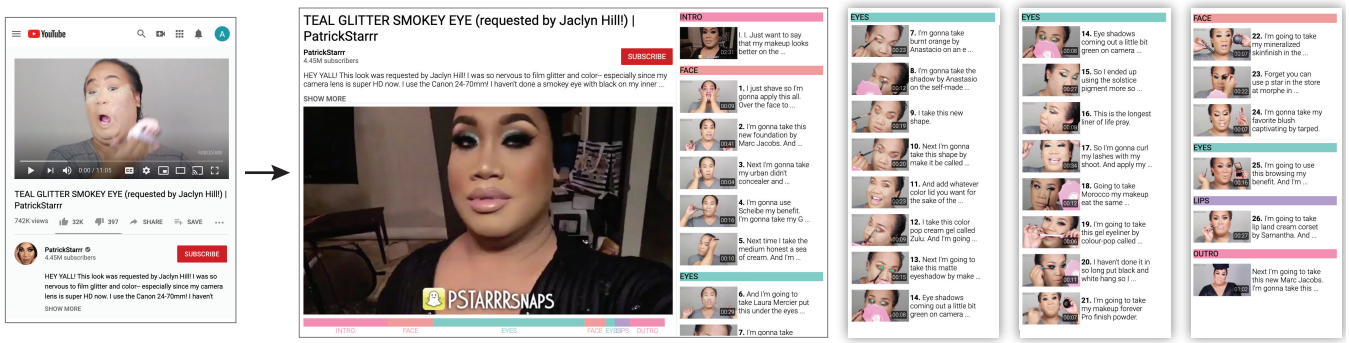
Anh Truong*
Stanford University
anhlt92@cs.stanford.edu

Peggy Chi
Google Research
peggychi@google.com

David Salesin
Google Research
salesin@google.com

Irfan Essa
Google Research, Georgia Tech
irfanessa@google.com

Maneesh Agrawala*
Stanford University
maneesh@cs.stanford.edu



(a) Input makeup video (b) Mixed media tutorial interface (c) Hierarchically grouped steps

Figure 1: Given an instructional makeup video as input (a), we automatically segment the video into a two-level hierarchical tutorial. We provide a mixed media UI that visualizes the resulting hierarchy (b). Viewers can use video, text, and images to navigate the instructions at both levels of the hierarchy. The hierarchy consists of fine-grain action steps organized into coarse-grained events based on the facial parts they manipulate (c). We encourage readers to zoom into the figure to see the text. Video Source: "TEAL GLITTER SMOKEY EYE (requested by Jaclyn Hill!)" by PatrickStarrr is licensed under CC BY.

ABSTRACT

We present a multi-modal approach for automatically generating hierarchical tutorials from instructional makeup videos. Our approach is inspired by prior research in cognitive psychology, which suggests that people mentally segment procedural tasks into event hierarchies, where coarse-grained events focus on objects while fine-grained events focus on actions. In the instructional makeup domain, we find that objects correspond to facial parts while fine-grained steps correspond to actions on those facial parts. Given an input instructional makeup video, we apply a set of heuristics that combine computer vision techniques with transcript text analysis to automatically identify the fine-level action steps and group these steps by facial part to form the coarse-level events. We provide a voice-enabled, mixed-media UI to visualize the resulting hierarchy and allow users to efficiently navigate the tutorial (e.g., skip ahead, return to previous steps) at their own pace. Users can navigate the hierarchy at both the facial-part and action-step levels using click-based interactions and voice commands. We demonstrate the effectiveness of segmentation algorithms and the resulting mixed-media UI on a variety of input makeup videos. A user study shows that users prefer following instructional makeup videos in our mixed-media format to the standard video UI and that they find our format much easier to navigate.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**.

KEYWORDS

Video Tutorials, Video Navigation, Video Segmentation.

ACM Reference Format:
Anh Truong, Peggy Chi, David Salesin, Irfan Essa, and Maneesh Agrawala. 2021. Automatic Generation of Two-Level Hierarchical Tutorials from Instructional Makeup Videos. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3411764.3445721

*This work was done while the first and the fifth authors were respectively an intern and an academic consultant at Google.

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05.
https://doi.org/10.1145/3411764.3445721

1 INTRODUCTION

Instructional videos help viewers accomplish or learn new tasks by walking them through the procedure in a step-by-step manner. As demand for remote learning grows, people are increasingly turning to online instructional videos as guides for completing everyday tasks and for learning new skills. For example, a 2017 study found that "how-to" videos earn the most attention of any content category on YouTube [29]. Instructional makeup videos, in particular, are extremely popular. A 2020 Google video search for "makeup tutorial" yields over 56 million hits, which collectively have billions of views. The top makeup videos amass millions of views with the most popular ones amassing over 205 million views [31].

While such instructional makeup videos are especially effective for conveying actions that are difficult to describe with words or static images, following them can be challenging. The linear nature of video forces viewers to watch the procedure at the pace of the video (which is often either too fast or too slow) [10, 27], or scrub a timeline in order to jump ahead or return to earlier steps. Scrubbing the timeline is especially difficult while applying makeup because both hands are often occupied in the makeup application task.

Prior work in cognitive psychology suggests that people mentally segment procedural tasks into event hierarchies, where coarse-grain events focus on objects while fine-grain events focus on actions [16, 42, 43]. The resulting hierarchical event structure aids in task comprehension as people understand the instructions more quickly when they are segmented in this way [43]. Although content creators often organize makeup videos to follow this hierarchical event structure, the videos typically present the content as a linear stream of information without marking even the fine-level action step boundaries.

In this work, we present a multi-modal approach for automatically segmenting an instructional makeup video into the two-level, object-action hierarchy and visualizing the resulting segments in the form of a hierarchical mixed-media makeup tutorial (Figure 1). We first analyze the domain-specific structure of makeup tutorials and find that they follow the object-action hierarchy suggested by cognitive psychology research with coarse-grain objects corresponding to facial parts and fine-grain steps corresponding to actions on those facial parts. We also find that each action step generally introduces a makeup product (e.g., eye shadow) and then applies it to some part of the face (e.g., the eyelid), sometimes using a specialized tool (e.g,. a brush). We leverage this structure to design a pipeline that combines computer vision techniques (e.g., face detection, object detection) and transcript text analysis (e.g., phrase detection, dependency parsing) to automatically identify action steps at the fine-level and then group together these steps based on the facial part at the coarse-level. We show that our multi-modal segmentation approach, specialized to the features of instructional makeup video, is more accurate than general-purpose shot detection or transcript-based segmentation methods techniques at the fine-grain level. Moreover, unlike previous fully-automated methods for segmenting instructional video that only provide fine-grain action steps [26, 44], our work is uniquely able to identify both the fine- and coarse-grain segmentation for makeup videos. Finally, we present a mixed-media navigation UI that visualizes the hierarchical structure of the makeup tutorial using text and video, and allows users to navigate the hierarchy at both the facial-part and action-step levels using click-based interactions and voice commands.

We demonstrate the effectiveness of our approach by generating 40 hierarchical makeup tutorials from input makeup videos authored by many different creators and covering a variety of different facial parts, application techniques, and styles. We compare these automatically segmented results against ground truth segmentations for 10 of these videos and find good agreement between the two. In an eight participant user study comparing our hierarchical mixed-media tutorial UI to the standard YouTube video UI, we find that by visualizing the two-level structure, our tutorials help users more easily identify and navigate to relevant sections of the tutorial — making it easier for users to follow the tutorial at their own pace and in the order of their choosing.

In summary, our work makes the following contributions:

(1) We conduct a formative analysis of existing instructional makeup videos and books to identify how they are hierarchically structured based on objects (facial parts) and actions (makeup application).
(2) We show how this two-level structure can be automatically extracted from an input video using a multi-modal segmentation approach leveraging a novel combination of computer vision and text analysis techniques.
(3) We show how the input video can then be transformed into a mixed-media tutorial that visualizes the two-level structure and facilitates click- and voice-based navigation.
(4) We validate the effectiveness of the resulting hierarchical design via a two-part user study.

2 RELATED WORK

We discuss three main areas of prior work that our work builds on: (1) analysis of instructional videos to extract steps, (2) extracting tutorial hierarchy, and (3) interactive tutorial navigation.

2.1 Extracting Steps from Instructional Videos

Researchers have proposed methods for automatically locating a sequence of steps in instructional videos of physical tasks, where a step is an event triggered by an action and often involves one or more objects. Alayrac et al. [2] and Sener et al. [33] leverage large datasets of narrated instructional videos for a simple everyday task that involves a small set of objects (e.g., cooking eggs or repotting a plant) to learn a sequence of common steps for that task and then localize those steps inside each individual video. Others have proposed automatic approaches for applying domain-specific knowledge (e.g. cooking knowledge), to align instructional videos (e.g. preparing a dish) with a text instructions (e.g. a text recipe) for the same task [26] or to a pre-specified list of steps [44]. More recent methods are able to identify visual correspondence between objects in a video frame and phrases in the transcript [17, 18]. We are inspired by these multi-modal techniques for extracting action steps from an instructional video. We similarly utilize both video frames and a transcript in order to identify the steps. However, this prior work focuses on either segmenting simple, general events or is specifically designed for cooking videos, and therefore cannot easily transfer to our domain of instructional makeup videos. Moreover, none of them are designed to perform coarse-grain segmentation.

2.2 Computational Tutorial Generation

Creating effective instructions can be a time-consuming process. Early research has developed automatic techniques for generating static instructions for furniture and toy assembly tasks [1, 16]. In the domain of software, a significant number of efforts have been designed to capture workflows from an expert's demonstration of a task (the sequence of operations they perform to complete the task) and convert them into useful tutorials. For example, some researchers have developed tools to simultaneously capture software operations and a screencast video, and convert them into a tutorial document with text instructions and annotated step images [14], supported by segmented video playback [7]. Meshflow [9] and Chronicle [15] build playback interfaces for mesh construction and graphical document editing workflows by automatically capturing and clustering the workflow operations. Others have focused on producing instructional videos for physical tasks in the real-world, by semi-automatically editing creator annotated raw video footage of a single-take demonstrating the task [8] or multiple annotated takes including b-roll [35]. While these prior techniques all generate structured output at the fine-granularity of individual action steps, they do not extract or present the object-action instruction hierarchy. In contrast our work focuses on revealing the hierarchy in pre-edited instructional makeup videos via *a fully automatic hierarchy extraction approach.*

2.3 Interactive Tutorial Controls

Tutorials are commonly presented as a video or a web document, which can be challenging to navigate and consume for tasks with high complexity [5, 7, 32, 36]. Researchers have proposed new interfaces for faster access to instructional content. Kim et al. [23] crowdsource step-by-step information from tutorial videos and visualize this information as thumbnails and text annotations on an interactive timeline. Chi et al. [7] present a mixed-media interface that combines text, images, and videos for photo manipulation tasks. Viewers can glance through a step-by-step document, locate a specific step, and review its video segment for detailed instructions. Fraser et al. [11] and Pavel et al. [30] use a chapter/section structure for videos of long duration to enable viewers to skim the content and replay a segment. To impose this chapter/section structure, Fraser et al. use heuristics specific to creative live streams while Pavel et al. focus on lecture videos and enable either the authors themselves or crowdworkers to construct this structure. Weir et al. [39] and Nawhal et al. [28] enable users to navigate tutorial videos by identifying sub-goals or milestones within the tutorial. Weir et al. use learner-sourcing to identify steps and subgoals for each video while Nawhal et al. focus on recipe tutorial videos and use a combination of manual annotation and computer vision techniques to extract structure for each video.

In our context of makeup instructions we draw on prior research in cognitive psychology, as well as observations of well-designed makeup instructions to find that the appropriate hierarchical structure divides the makeup application process into coarse-grain facial parts and fine-grain actions steps. We develop a fully automated approach for extracting this hierarchical structure from an input makeup video. Based on this two-level hierarchy, we design a mixed-media tutorial interface that aids users in navigating the tutorial to skip sections and re-watch specific steps.

3 STRUCTURE OF INSTRUCTIONAL MAKEUP VIDEOS

Cognitive psychology research suggests that people mentally consider procedural tasks as event hierarchies, where coarse-grain events focus on objects while fine-grain events focus on actions that manipulate those objects [16, 42, 43]. Traditional static tutorials as found in books often follow this two-level hierarchy. For example, a recipe for cooking a Thanksgiving turkey may be coarsely organized based on the objects (or items) being prepared — e.g., make the *stuffing*, roast the *turkey*, and cook the *gravy*. Then, within the "roast the turkey" event, fine-grain action steps might include: "1. Take giblets out of turkey," "2. Place turkey in large roasting pan," "3. Salt and pepper inside of turkey," and so on. The resulting hierarchical event structure aids in task comprehension as people both understand the instructions more quickly when they are segmented leveraging the structure [43].

In order to understand the event structure in the instruction makeup domain, we examined 50 instructional makeup videos from 20 different YouTube creators as well as two instructional makeup books [3, 20]. We focused on videos in which only a single person appears on screen as these are most common for makeup videos. We found that all of these videos follow the object-action hierarchy suggested by cognitive psychology research with coarse-grain objects corresponding to facial parts and fine-grain steps corresponding to actions on those facial parts. For example, rather than switching back and forth between applying eye makeup and contouring the face, creators focus on one part of the face at a time; they finish applying makeup to one part of the face before moving on to another part.

Similarly, makeup books are divided into chapters based on the parts of the face (i.e., face, eyes, lips), and each chapter then contains instructions for makeup techniques that manipulate that part of the face (e.g., eyeliner application in the eyes section). These instructional makeup videos and books rarely organize steps by tools or products. Instead, tools and products are often used to search for specific steps.

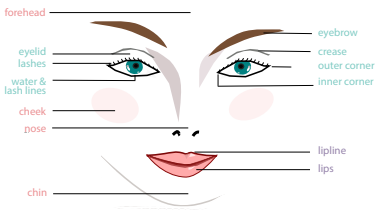


Figure 3: Components of the face on which makeup is typically applied. Components are color coded by facial part category (teal for eyes, pink for face, purple for lips)

Across all the instructional makeup videos and books, the coarse-level events focus on three categories of facial parts (Figure 3):

- ***Lips.*** Encompass the lip line and the lips themselves.
- ***Eyes.*** Encompass the eyebrows, lashes, eyelids, crease, inner and outer corner of the eyes, water lines, lash lines, and the area under the eye.
- ***Face.*** Encompass the skin, cheeks, nose, chin, and forehead.

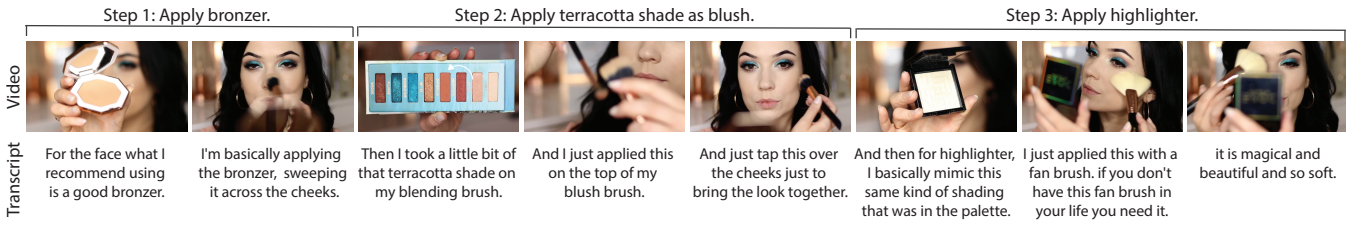


Figure 2: Frames from an instructional makeup video (Video B in Table 2). In a series of three action steps, the creator first applies bronzer (step 1), then applies a terracotta shade (step 2) and finally applies the highlighter (step 3). All three steps start by introducing a makeup product (bronzer, terracotta shade in the color palette, highlighter). The creator continues steps 1 and 2 by showing shots of her applying the products with a brush as she describes those makeup application actions in the narration. In step 3, she also shows two shots of her applying the makeup with a brush tool. However, in the narration, she only describes the application of the makeup in a quick phrase and spends the rest of that step commenting on her opinions of the brush. *Image source: Glam Makeup Tutorial by TheMakeupChair under CC BY.*

Within a coarse-level event, we found that each fine-grain action step consists of a sequence of visuals demonstrating a makeup product being applied to the face, and narration (or text in the case of books) describing how to execute that action (Figure 2). Take the video tutorial shown in Figure 2 as an example. First, the creator introduces the product they will use in the step (Figure 2 all three steps). Next, if they are using a tool to apply the product, they may introduce this tool (Figure 2 step 3). Otherwise, they directly move on to demonstrating and/or describing how to apply the product (Figure 2 all three steps). They may stop narrating or give some additional commentary as they apply the makeup (Figure 2 step 3). In makeup videos, the visual sequence and narrative sequence are not always perfectly aligned as the creator may choose to start the narration before the visuals appear or introduce visuals for a new step before they finish commentary about the current step.

In addition to the instructional content, all the makeup videos that we examined consisted of two non-instructional sections, an introduction and a conclusion, at the start and end of the video respectively. Creators use the introduction section to introduce themselves, describe the look and its motivation, occasionally introduce products that they will focus on in the tutorial, and ask viewers to subscribe to their channel. The conclusion section allows the author to show off the finished look, once again ask viewers to subscribe, show bloopers, and point to their other videos.

4 ALGORITHMIC APPROACH

Given an instructional makeup video, our goal is to algorithmically extract the fine-grain action steps and group these steps based on facial parts to form the coarse-level events. Whereas prior work in computer vision and NLP has often operated at a low-level to identify objects or sub-actions independently on instructional video or text, a key feature of our approach is that it combines such low-level visual and textual information to more robustly segment makeup videos into the two level hierarchy.

We start by using Google Cloud's Speech-to-Text API [13] to obtain a time-aligned transcript of the input video. Our multimodal approach then works in three phases (Figure 4). In phase 1, oversegmentation-and-labeling, we oversegment the video into shots using shot detection techniques, and we break the transcript into spoken phrases using punctuation detection techniques. We then label facial parts, makeup products, tools, and makeup application actions in each resulting shot or phrase. Since the shots and phrases sometimes contain complementary label information, we construct shot-phrase pairs that allow us to consider these labels jointly. In phase 2, we construct the fine-grain action steps, by grouping the oversegmented shot-phrase pairs that are part of the same makeup application step. Our approach is to search for a pattern of product introduction followed by makeup application within the labeled shot-phrase pairs to determine step boundaries. Finally, in phase 3 we construct the coarse-grain facial-part groupings by clustering sequences of action steps that apply to the same part of the face.

As noted in Section 3, instructional makeup videos usually contain introduction and conclusion sections. The introduction consists primarily of commentary, and may include some product introductions. Similarly the conclusion is mostly comprised of commentary but may also reiterate the products used in the instructions to remind viewers what they need to buy to replicate the look. Once we have extracted the action steps, we segment the portion of the video that appears *before* the first action step as the introduction segment and the portion that appears *after* the final action step as the conclusion segment.

4.1 Phase 1: Oversegmentation and Labeling

In phase 1, we break the input video and transcript text into very short pieces (e.g. shots and phrases) that are smaller than the fine-grain action steps, and we label these pieces as *product introductions*, *tool introductions*, *makeup application*, or *commentary*. We pair the shots and phrases based on temporal overlap which allows phases 2 and 3 of the algorithm to reason about the video and text together.

4.1.1 Video Shot Detection. In order to segment the video into visual shots, we apply the edge change detection approach of Lienhart et al. [25] to locate shot boundaries. Our algorithm accepts a frame as a boundary if the edge change ratio between this frame and the previous frame is greater than a user defined threshold ϵ, which we empirically set to 0.4. We tested all thresholds in our pipeline on multiple makeup videos and chose constants that produced consistently good results. The resulting shots are 0.2 to 34 seconds long for the makeup videos we tested.

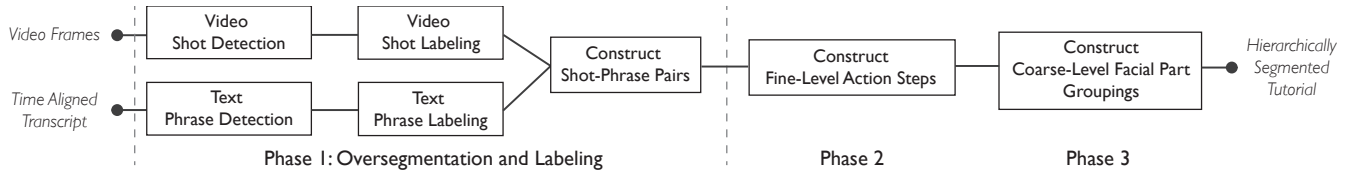
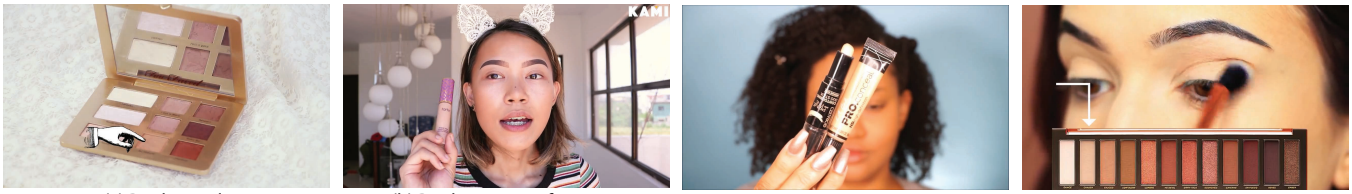


Figure 4: Our algorithmic pipeline consists of three phases. Given an instructional makeup video and a time-aligned transcript as input, Phase 1 oversegments and labels the video and transcript to form shot-phrase pairs. Phase 2 constructs fine-grain action steps by grouping together the shot-phrase pairs that are part of the same makeup application action step. Phase 3 constructs the coarse-grain facial part groupings by clustering sequences of action steps that apply to the same facial part.



(a) Product only (b) Product next to face (c) Product still and in front of face (d) Product overlaid on face

Figure 5: Product introduction frames emphasize the product by only showing the product (a), placing the product next to the creator's face onscreen while they are introducing it (b), or placing it over the creator's face, but keeping it relatively fixed in place (c,d). In (d) the product (eye shadow) is overlaid on the face to introduce it and the creator is simultaneously applying it in the background (d). *Image sources: (a) 1920s Wearable Makeup Tutorial by Jbunzie, (b) Easy Glowy Summer Makeup by BeKami, (c) Morphe Brushes 35R Palette Makeup by Brittney Enora (d) Smokey Eye Makeup Look + Face & Lips by TheMakeupChair. All licensed under CC BY.*

4.1.2 Video Shot Labeling. For each resulting shot, we first construct low-level labels identifying facial parts, objects, and motions within the shot. We then use these low-level labels to construct higher-level labels for product introduction, makeup application, and commentary. To construct the low-level labels, we sample one frame per second and then label the frame using computer vision techniques. We run the MediaPipe Facemesh pipeline [41] to obtain 486 3D facial landmarks per face in the frame. We use these landmarks to construct bounding boxes around the eyes, lips, and face and use the resulting labels for facial-part-grouping in phase 3 of our algorithm. We also apply the object detection approach of Huang et al. [19] to detect and locate products and tools in each frame. This detector outputs a list of bounding boxes with an object category label and confidence score. We have found that its "packaged goods" label encapsulates most of the products and tools used in makeup tutorials. Thus, we retain all resulting objects labeled "packaged goods" with a confidence score ≥ 0.5. Finally, to differentiate whether a makeup product is in use or just being introduced, we estimate the motion taking place within each frame using the frame-differencing approach of Kameda et al. [21] which computes the difference of each sample frame with its two immediate neighboring frames. The output of the motion estimation algorithm is a binary pixel intensity map where the black pixels are static and the white pixels indicate movement. We use these low-level labels to label each sample frame within the shot as a product introduction, makeup application, or commentary.

Labeling product introduction frames. During a product introduction the makeup video typically focuses on the product by emphasizing it in the frame (Figure 5), rather than the person using the product. Thus, we label a frame as a product introduction if it contains a product and the product is relatively still. Specifically, we calculate a product's movement as the ratio of motion pixels to total pixels in the product bounding box and check whether this product movement ratio falls below m, set empirically to 0.05.

(a) Using product (lip liner) (b) Using tool (brush) (c) Using hands

Figure 6: Makeup application frames usually involve significant motion of a makeup product (a), a tool (b) or the creator's hands (c) over the face. *Image sources: (a) Smokey Eye Makeup Look + Face & Lips by TheMakeupChair (b) How to Apply Flawless Foundation by Brittney Enora, (c) Yellow Cut Crease Makeup by PatrickStarrr. All licensed under CC BY.*

Labeling makeup application frames. During makeup application the creator is usually moving the product, a tool or their hands in front of their face (Figure 6). Thus, we mark a frame as a makeup application frame if it contains a face with significant movement over it – i.e. the movement over the face bounding box exceeds $m = 0.05$. Note that when calculating this movement we exclude the area around the mouth and eyes since these facial parts tend to move a lot when the creator talks or blinks. Occasionally, the creator may choose to both introduce a product and demonstrate its application in the same frame (Figure 5d). Our algorithm separately calculates movement over both the product and the face to detect

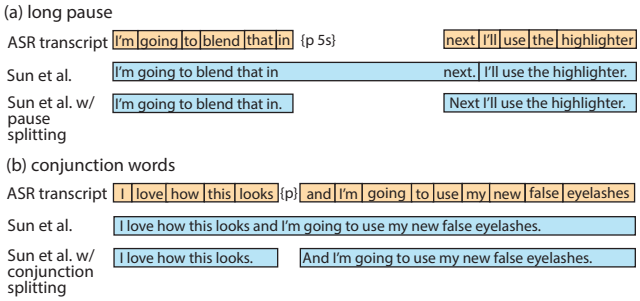


Figure 7: Makeup video creators often pause (denoted by {p}) between phrases (a) and/or join phrases using conjunction words (b). Sun et al.'s [34] punctuation model does not account for these aspects of spoken narration. Extending their approach with our pause-based (a) and conjunction-based (b) splitting correctly splits the transcript.

a) Makeup Products Nouns

balm	concealer	filler	lipstick	powder	spray
blush	contour	foundation	mascara	primer	stain
bronzer	cream	gloss	moisturizer	remover	
brow gel	curler	highlighter	palette	setting	
color	eyeliner	lashes	pencil	shade	
color correcter	eyeshadow	lip liner	plumper	shine control	

b) Makeup Tool Nouns

applicator	blender	brush	finger	sponge	tweezer

c) Makeup Application Verbs

add	buff	curl	highlight	put	stick
apply	build	draw	line	set	stack
bake	clean	fill	pat	smooth	stroke
blend	conceal	finish	pinch	smudge	tap
brush	contour	flick	prime	spray	

Table 1: Common types of makeup products (a), tools (b), and verbs that specify a makeup application action (c). We obtain these lists by examining a variety of instructional makeup videos and shopping hierarchies for beauty product retailers.

both the product introduction and makeup application labels for this frame.

Labeling commentary frames. During commentary the creator can be onscreen speaking directly to the camera, but without applying makeup or showing a product, or the creator may be narrating from off screen without any product being shown. Thus, we mark a frame as commentary if it does not contain a product or tool and either contains a face with low movement (e.g. movement less than $m = 0.05$) or does not contain a face.

4.1.3 Text Phrase Detection. Makeup video creators usually narrate their video using a series of short phrases with pauses and conjunctions between them, but without ending a complete sentence. The automatically generated transcript text associated with a video is a continuous stream of words with timestamps. Typically, it does not contain punctuation (e.g., periods and commas) that delimit phrases. Therefore, we extend the text punctuation model of Sun et al. [34], which was designed to split written text into sentences, to segment the transcript text into phrases. Specifically, we extend their model to consider (1) the timing of pauses between words and (2) conjunction words that mark phrase boundaries.

Split based on pause timing. The phrase-level grouping of spoken words depends on their temporal position. For example, in Figure 7a, the creator takes a long five-second pause between two phrases to execute the action in the first phrase. Without the context of the pause, Sun et al.'s algorithm groups the word "next" with the first phrase. To ameliorate this issue, we first split the transcript into groups of words based on the duration of pauses between them. If the pause between two words lasts for more than δ seconds, our algorithm creates a new word group starting at the second word. In practice we have found that setting $\delta = 1.0$ works well across a range of makeup videos made by a variety of creators with different speaking styles. After splitting the transcript based on pause duration in this manner, we apply Sun et al.'s punctuation model to each resulting word grouping. We treat each resulting comma or period as a phrase boundary.

Split based on conjunction words. Even after grouping the words based on pauses, the segmentation produced by Sun et al.'s can be incorrect because creators often do not speak using proper sentences. Instead, they regularly connect a series of phrases using conjunction words such as "and," or "so." For example, in Figure 7b, the creator comments on the result of one step and then jumps to a description of the next step using the word "and" to join the phrases. Sun et al.'s punctuation model combines the two phrases into one conjoined sentence. To account for these under-segmentations we use Google Cloud's Natural Language API [12] to apply part-of-speech tagging, dependency parsing, and lemmatization to all the sentences generated by Sun et al.'s model. If a sentence contains a conjunction word (e.g., "and," and "so,") and that conjunction relates two verbs, we split the sentence into phrases at the conjunction. We also retain the part-of-speech tags for later use.

4.1.4 Text Phrase Labeling. We label each resulting transcript phrase as a *product introduction* phrase, a *tool introduction* phrase, a *makeup application* phrase or a *commentary* phrase using a template matching approach. For example, the template for labeling a product introduction looks for phrases that contain an *introductory expression* (e.g. "take", "use", "pick up", "this is the") and a noun from a list of *makeup product nouns* (Table 1a). Similarly the template for labeling a tool introduction looks for an *introductory expression* and a noun in a list of *makeup tool nouns* (Table 1b). Our template for the makeup application label simply looks for a makeup application verb (Table 1c) and if none of these templates apply to a phrase, we label the phrase as commentary. A phrase may match more than one of these templates. For example, the phrase "using our flat brush, we are going to apply the eye shadow onto the crease" contains both a tool introduction (i.e., the flat brush) and a makeup application (i.e., apply the eye shadow). We label the phrase with all the templates it matches.

We manually constructed these tables of template-matching expressions and words based on close examination of a variety of instructional makeup videos as well as shopping sites for beauty product retailers. Occasionally, creators will only reference products by their official branding name, which may not explicitly state the product type. To account for these cases, our algorithm tags the phrase with a product introduction if the accompanying noun does

not occur in either the product or tool lists. Our algorithm also keeps track of all product and tool types mentioned in the phrase. Finally, we check if the phrase contains a reference to a facial part using the list of facial part words in Figure 3. We use these labels in phase 3 of our algorithm to group steps by facial parts.

4.1.5 Construct Shot-Phrase Pairs. The video and the transcript often contain complementary information. For example, the shot may be labeled to contain a makeup product introduction as the frame shows eye shadow, while the narration may be labeled as makeup application because it simply says "apply this [while pointing to eyeshadow] to the crease". By leveraging the labels for visual shots and text phrases labels together, we develop more effective heuristics for segmenting the video. Our approach constructs shot-phrase pairs between temporally overlapping shots and phrases.

Given a visual shot v with duration $|v|$ and a spoken phrase t with duration $|t|$, we first compute the maximum overlap **O** between them as

$$\mathbf{O}(v,t) = \max\left(\frac{|t \cap v|}{|t|}.\frac{|t \cap v|}{|v|}\right). \quad (1)$$

If the overlap is greater than a threshold ω, we associate v and t as a shot-phrase pair. We empirically set the threshold to avoid constructing pairs when there is a relatively small amount of overlap between the shot and phrase. For example, makeup videos sometimes use a J-cut where the narration for the next step starts a little before a shot change visually depicts that step. In such cases the spoken phrase should only be paired with the shot that starts later. We find that setting $\omega = 0.3$ works well across a variety of instructional makeup videos.

Note that a phrase can belong to more than one shot-phrase pair; for example if a phrase significantly overlaps with two adjacent shots it will form a pair with both of them. Similarly, a shot can belong to more than one shot-phrase pair; for example if multiple phrases are spoken over the shot, a shot-phrase pair is formed for the shot and each such phrase. In addition, a shot-phrase pair may include an empty phrase if there is no narration during the shot. Thus, each shot-phrase pair contains exactly one visual shot and zero or one phrases.

4.2 Phase 2: Construct Action Steps

We next group together shot-phrase pairs to form the fine-grain action steps. Our approach is to look for a pattern makeup video creators commonly use when demonstrating a step. As noted in Section 3, creators usually begin an action step by introducing the product they will use, sometimes along with a tool they will use to apply it. They then demonstrate how to apply the product to the face. They may either stop narrating or provide additional commentary as they apply the makeup. Therefore, we iterate through the list of shot-phrase pairs looking for a *product introduction pair* and mark such a pair as the start of a candidate action step. We then group subsequent shot-phrase pairs into the candidate action step until we encounter the next product introduction pair. To determine whether a shot-phrase pair is a product introduction, we examine the following conditions, if:

(1) the phrase is labeled as a product introduction and the product noun or name is not the same as the product noun or name in the previous action step, or
(2) the visual shot contains a frame labeled as a product introduction frame and the bounding box of the product is visually dissimilar from the product depicted in the previous action step (i.e. a visual difference of ≥ 0.2 measured using use Wang et al.'s [38] deep ranking image similarity model across the entire frame).

In practice, we have found that the product introduction label on visual shots cannot always reliably differentiate between product introductions and makeup applications. Therefore, we further check that any shot-phrase pair that passes the second condition is not also labeled as makeup application.

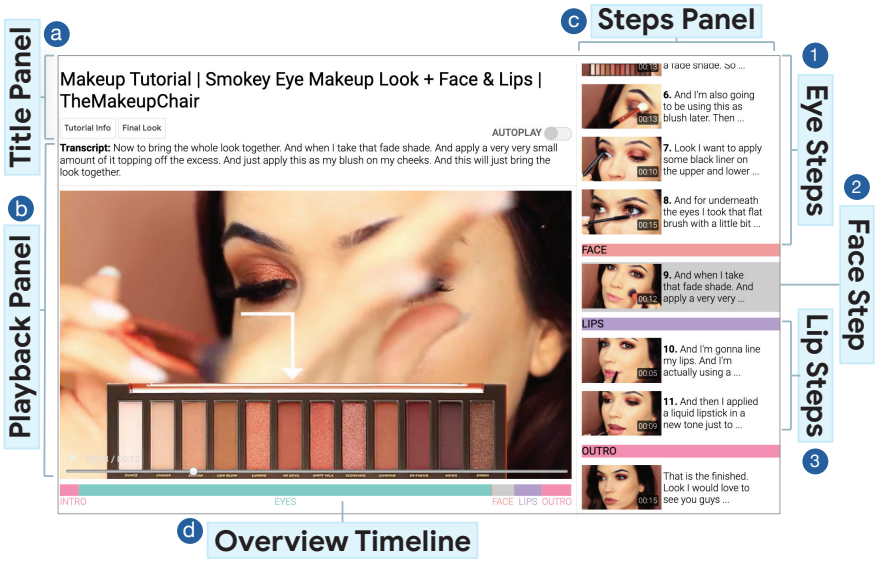
Once we have a set of candidate action steps we further check that each such candidate contains a shot-phrase pair we can identify as a *makeup application pair*. For each shot-phrase pair within the candidate step we check if the phrase is labeled as makeup application or the visual shot contains a majority of frames labeled as makeup application. If the candidate action step does not contain a makeup application pair we remove it from our list of action steps.

4.3 Phase 3: Construct Facial Part Groupings

To form the coarse-grain facial-part groupings, we first consider each action step and compute the facial part it most likely to focuses on. Specifically, we count the number of times each facial part appears as a label on the shots and phrases within a step and treat the label with the highest count as the facial part for the step. If the step does not contain any facial part labels (i.e., the creator does not explicitly mention the facial part they are manipulating and our face detector cannot detect the part due to occlusion or motion), we assign it the facial part label of the previous step. We then cluster all contiguous action steps which share the same facial part label into one facial part group. As a final cleanup step, we merge any facial part group that contains only a single action step with its neighbors if both neighboring facial part groups act on the same facial part. Creators are unlikely to switch from doing their lip makeup, for example, to work on their eye makeup for one step before returning to the lips. In practice we have found that these lone segment groups can occur when a creator is applying makeup to one facial part but refers to another facial part for comparison or commentary.

5 NAVIGATION INTERFACE

Figure 8 shows our hierarchical mixed-media interface. The *title panel* presents the title and creator information of this video tutorial (Figure 8a). Viewers can browse the steps of the tutorial by scrolling through the *steps panel* (Figure 8c). In this panel, the fine-grain makeup application steps are grouped by facial part (Figure 8c 1-3) so that viewers can directly see the two-level hierarchy. Each step includes a thumbnail and summary text that describes the step. Viewers can hover over the step thumbnail to preview the video for the step. They can click a step to play back the step video in the *playback panel* (Figure 8b), which also allow them to play, pause and scrub through the video segments of the current step. Viewers can click on one of the *overview timeline* bars to play back the steps for that section (see Figure 8d). The overview timeline bars and section headers in the step panel are color coded by facial part (e.g., teal



Voice Commands
Play the (section name) section
(Next | Previous) section
Show me the (face sub-component | tool | product)
Go to step (step number)
(Next | Previous) step
Replay Step
Go (forward | back) (number) (seconds | minutes)
Play
Pause

Figure 8: On the left is a mixed-media video tutorial automatically generated by our pipeline. It contains four main components: (a) the title panel displays tutorial title, creator name, description, and final look; (b) the playback panel enables the user play back and scrub through a single step; (c) the steps panel lets the viewer navigate the tutorial using finer-grain actions; and (d) the overview timeline visualizes the length and order of the tutorial's coarser-grain facial part groupings. Currently the viewer is playing step 9. On the right is the full set of voice commands for the hierarchical mixed media interface. Users can navigate at both the coarse-grain level by asking to jump to specific *sections* and at the fine-grain level by asking to jump to specific steps, as well as using specific facial sub-components, tools, products or time. *Video source: Smokey Eye Makeup Look + Face & Lips by TheMakeupChair under CC BY.*

for eyes, salmon for face and purple for lips) to further emphasize the coarse-level of the hierarchy.

The transcript view above the video shows the fully transcribed narration for the current step. The viewer can click on a word in the transcript to jump to the time in the video when that word is spoken. The currently focused step (such as step 9 in Figure 8) is highlighted both in the steps panel and the overview timeline. By default, our UI continuously plays the steps during video playback. Viewers can choose to playback one step at a time via the auto-play toggle. These functionalities enable viewers to follow the tutorial sequentially or jump around as they please.

For each step, we automatically generate the step thumbnail previews by sampling the video corresponding to the step at a rate of 7 frames per second in the animated GIF format. To help viewers focus on the content that is the most relevant for execution in a text summary, we automatically concatenate all transcript phrases within the step that are labeled as either product introduction or makeup application and skip phrases marked as commentary. Usually the step text is comprised of a few phrases. We use ellipses to indicate when the step text is longer than the space available in the steps panel.

Voice-based navigation UI. Often, applying makeup requires both of the user's hands, making it difficult for them to control the interface with click or touch-based interactions [37]. We therefore provide voice-based navigation controls so that users can vocally skip to a different part of the tutorial. Recent studies have examined voice-based UIs for navigating instructional videos [5]. We adapted their findings and implemented voice commands that enable users to navigate the tutorial using five types of referents: facial parts, action steps, tools, products and time. Users can navigate at the coarse level by asking the interface to jump to a *section* describing a different facial part category, such as the eyes, by saying "go the eyes section". They can also jump to a specific facial part sub-component, such as the eyelashes, by saying "show me the eyelashes." In some cases, these commands may match with more than one step or section. The system shows the closest result after the current step first and the user can traverse the set of matches by saying "next result" or "previous result." At the fine-grain step level, users can jump between steps by saying "next step" and "previous step", or to a specific step by saying "go to step X", where X is the step number. They can also jump to a specific product or tool, such as the mascara, by saying "show me the mascara." The full list of commands is available in Voice Commands table in Figure 8.

6 RESULTS

To demonstrate the effectiveness of our pipeline, we generate hierarchical tutorials for 40 instructional makeup videos retrieved from YouTube (please find the full list in our supplementary materials). We collected ground truth segmentations for ten of these videos (Table 2). We selected single-person videos that capture a diversity of creators, editing styles, makeup looks, and lengths. Figures 1 and 9 shows seven of the resulting mixed-media tutorials Below we describe our observations.

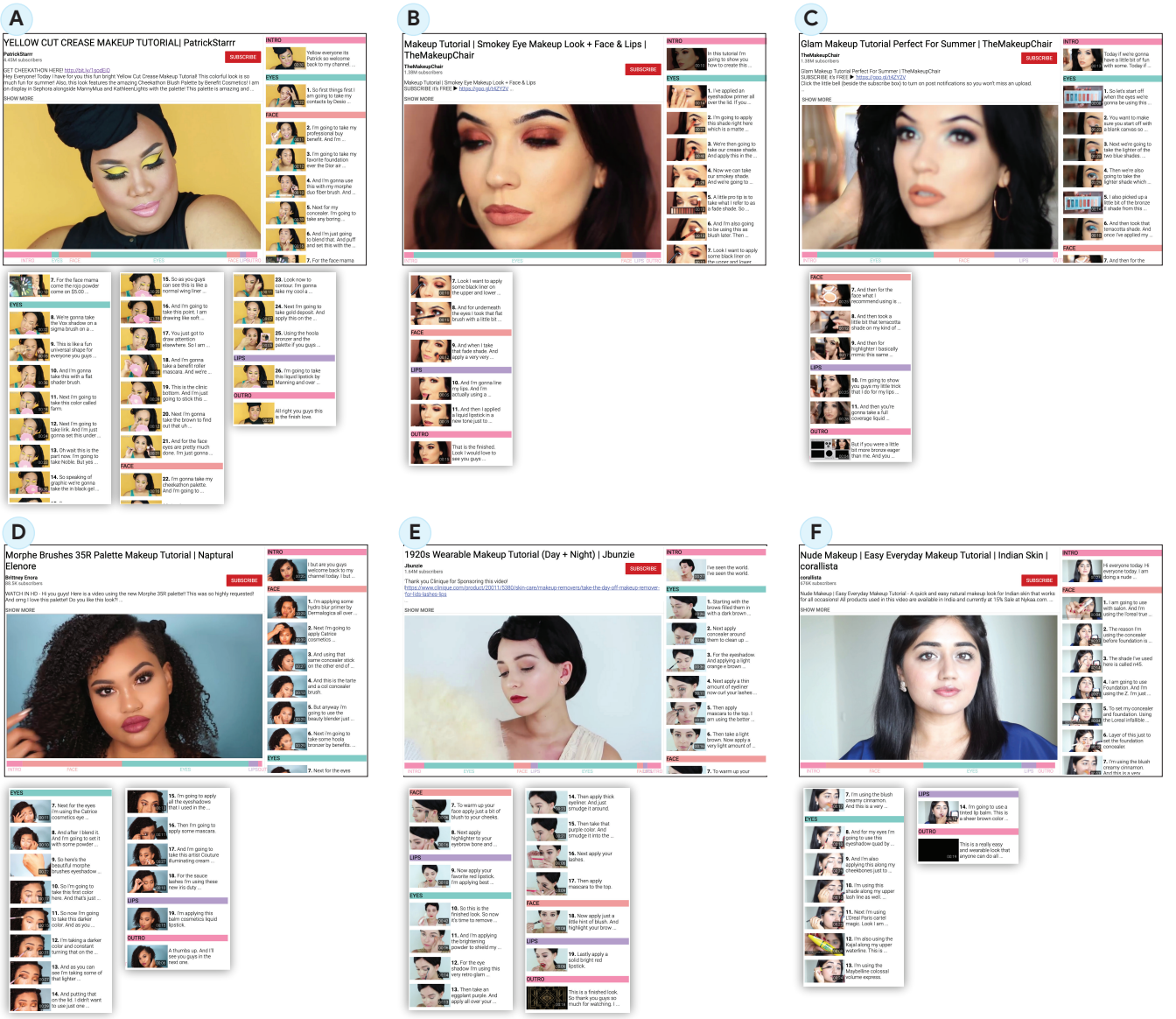


Figure 9: Mixed media makeup tutorials automatically generated from instructional makeup videos. The first six rows in Table 2 correspond to examples A-F here. For each example (A-F) we show the initial screen of the resulting tutorial in our interface (top) and the remaining steps in the steps panel (below). Each tutorial is segmented at the coarse-level by facial part (labeled with colored bars for EYES in teal, FACE in pink, LIPS in purple) and at the finer-level by action steps in which a makeup product is applied to that facial part (numbered steps). We encourage readers to zoom into the image to examine the step thumbnails and step text. *Video sources: (a) Yellow Cut Crease Makeup by PatrickStarrr, (b) Smokey Eye Makeup Look + Face & Lips by TheMakeupChair, (c) Glam Makeup Tutorial by TheMakeupChair, (d) Morphe Brushes 35R Palette Makeup by Brittney Enora, (e) 1920s Wearable Makeup Tutorial by Jbunzie, (f) Nude Makeup | Easy Everyday Makeup Tutorial by corallista. All licensed under CC BY.*

The coarse-level segmentation by facial parts (eyes, face or lips) is sensible as shown by the step thumbnails and step text. In all of these examples, the creators start by applying makeup to either the face (Figure 9A, D, F) or the eyes (Figure 9B, C, E) and work on the lips towards the end of the tutorial. In example A, the creator worked on the face in two parts – first prepping the face and then revisiting it later to contour. The step thumbnails and step text usually depict or mention the face part grouping they belong to. We observed a few coarse-level misgroupings. Occasionally a creator mentions two different parts of the face in the same action step and the step is grouped incorrectly. In example A step 21 for instance the creator comments on the state of their eye makeup while introducing face

makeup and the step is grouped with the eyes rather than the face. Such coarse-level misgroupings are relatively uncommon.

The fine-level segmentation by action steps is also sensible for these examples, though a small number of steps are oversegmented compared to the ground truth, breaking up a single action into multiple steps. In Figure 9A, steps 3 and 4 should be a single step explaining how to apply foundation while steps 13, 14 and 15 all about applying eye liner. Similarly in Figure 9B, steps 5 and 6 should be combined as they both are about applying a particular shade of eyeshadow, while in Figure 9F, steps 1 and 2 are both about applying concealer and steps 12 and 13 are both about applying eyeliner. While such oversegmentation increases the number of steps visualized in the mixed-media tutorial, because our interface automatically plays the next step the oversegmentation has a relatively small impact on usability. Occasionally steps are undersegmented and a single step describes application of more than one product. In Figure 9E, step 4 the creator is both applying eyeliner and curling the lashes in a single step. Unlike oversegmentation, such undersegmentation can hurt usability as it can be more difficult for viewers to navigate to actions they are most interested in following. However, participants in our user study were able to use the clickable transcript to easily locate the boundary between product applications in an undersegmented step.

6.1 Comparison to Ground Truth

To further gauge the accuracy of our two-level object-action event segmentations, we compare our automatically generated segmentation to manually generated ground truth segmentations. To generate the ground truth, we had one author and 10 non-author paid raters independently mark fine-level action steps and coarse-level facial parts for 10 of our input makeup videos. Each video was labeled by 3 different raters. All raters have previous experience watching makeup tutorials and applying makeup. Raters had access to a time aligned text transcript for easier navigation during the labeling process. On average, raters spent 3.25 minutes to label a one-minute video segment.

As shown in Table 2 the number of steps and the number of facial parts differ slightly between our algorithm and the three ground truth annotations. We compare two sets of segments, such as our algorithmic event segmentation to the ground truth, using an $F1$ score [40], which a standard measure of the difference between two sequences of segments that properly accounts for the relative amount of overlap between corresponding segments. We compute the average $F1$ score between two sets of segments as follows. Consider a pair of fine-grain action step segments,$(s_{\text{alg}}, s_{\text{gt}})$ where s_{alg} is generated by our algorithm and s_{gt} is from the ground truth. We can compute

$$F1(s_{\text{alg}}, s_{\text{gt}}) = \frac{2|s_{\text{alg}} \cap s_{\text{gt}}|}{|s_{\text{alg}}| + |s_{\text{gt}}|}, \tag{2}$$

where $|s|$ denotes the duration of the segment. In many instances, the raters marked slightly different (4 seconds or less) start and end boundaries for the same step. To avoid penalizing these small differences, we round all start and end times to the nearest 4 second boundary before computing F1. The $F1$ score has a range of 0 to 1. A higher score means greater overlap between the segments. As an aggregate measure of overlap, we average the $F1$ scores across pairs of segments that contain some overlap (i.e. $|s_{\text{alg}} \cap s_{\text{gt}}| > 0$). We then average the overlap measures across the three raters for each video. At the fine-grain level of action steps, the average $F1$ comparison ranges between 0.64 and 0.90 with an overall average of 0.80 (Table 2 col 7). At the coarse-grain level of facial parts, the average $F1$ comparison ranges between 0.50 and 1.0, with an overall average of 0.81 (Table 2 col 10). We note that these scores are relatively high and indicate good agreement between our algorithmically generated event segmentation and the ground truth.

In addition, we calculated the inter-rater reliability between the three raters for each tutorial as the $F1$ scores between each pair raters for the same tutorial. The average inter-rater reliability across all videos is 0.91 with a range of 0.81 to 1.0 at the fine-grain level and 0.93 with a range of 0.70 to 1.0 at the coarse-grain level. These scores demonstrate good agreement between raters.

6.2 Comparison to Ablated Pipelines

To evaluate the effectiveness of various components our fine-grain segmentation algorithm we compare the segmentations produced by our full pipeline to pipelines in which we have ablated certain components of phase 1 (Table 3). Relying only on video shot detection as in Pipeline (1), or only on text phrase detection as in Pipeline (2), yields poor average $F1$ of 0.37 and 0.24 respectively when compared to ground truth segmentations. Because creators often use multiple shots and speak more than one phrase in a single step, these basic shot and phrase detection pipelines tend to greatly oversegment the action steps and thus perform poorly. Pipeline (3), greatly improves the average $F1$ score to 0.54 because it combines information from both the video and the text transcript as it generates the fine-grain action steps. Pipeline (4), further improves the average $F1$ score to 0.61 by making use of the video shot labels to create a new action step each time a new product is shown in the video shot. Pipeline (5) similarly improves on the simpler strategy of Pipeline (3) by making use of the text phrase labels to create a new action step each time a new product is introduced in the voice over transcript. Pipeline (5) yields an average $F1$ of 0.77, which is better than the $F1$ score of Pipeline (4) indicating that the text phrase labeling provides better information for action step segmentation than video shot labeling. Finally, our full pipeline, Pipeline (6) yields an average $F1$ of .79; a modest improvement over Pipeline (5). It does this by combining information from the text phrase labels and video shot labels to form the action steps.

7 USER EVALUATION

We conducted a two-part observational user study. The goal of Study I was to evaluate how well users could follow makeup tutorials using our hierarchical, mixed-media interface compared to using the standard YouTube video playback interface. In Study II, our goal was to better understand how the hierarchical structure of our tutorials helped users comprehend them. To focus on the visual interface in a conventional video playback setting, we did not introduce the voice commands in these studies.

				Fine Grained					Coarse Grained		
Title (Duration)	**Creator**	**# Shots**	**# Phrases**	**# Steps**	**Avg. GT # Steps**	**# Steps Avg. F1**	**# Under Seg.**	**# Over Seg.**	**# Face**	**Avg. GT # Face**	**# Faces Avg. F1**
(A) Yellow Cut Crease Makeup (807s)	PatrickStarrr	211	256	26	25.3	0.80	4	4	5	5	0.75
(B) Smokey Eye Makeup (279s)	TheMakeupChair	65	66	11	12	0.88	2	1	3	3	1.00
(C) Glam Makeup (234s)	TheMakeupChair	64	73	11	11.33	0.79	2	2.33	3	3	0.69
(D) Morphe Brushes 35R Palette (423s)	Naptural Elenore	75	91	19	22	0.77	4.67	3	3	3	0.84
(E) 1920s Wearable Makeup (358s)	Jbunzie	76	58	19	23	0.90	3	0.33	6	8	0.88
(F) Nude Makeup (272s)	corallista	82	82	14	10.33	0.73	1.33	4.33	3	3	1.00
(G) Day to Night Makeup (492s)	corallista	95	121	30	18	0.64	2.33	10.00	5	6.33	0.50
(H) How to Apply Flawless Foundation (778s)	Brittney Enora	110	153	18	16.33	0.82	2.33	4	1	1	1.00
(I) Teal Glitter Smokey Eye (666s)	ParticKStarrr	159	196	26	26	0.80	5.33	4	5	5	0.80
(J) Easy Glowy Summer Makeup (321s)	BeKami	74	60	14	13.67	0.85	2	2.67	4	5	0.66

Table 2: We evaluated our approach on ten instructional makeup videos that cover a variety of durations, creators, makeup looks, and editing styles. For each tutorial we report the number of shots (col 3) and phrases (col 4) generated in phase 1 of our algorithm. We report the number of fine-grain action steps (col 5) it generates in phase 3 and the number of coarse-grain facial parts groupings it identifies (col 10) in phase 4. For comparison, we report the average ground-truth number of action steps (col 6) and number of facial parts groupings (col 11) as well as the average F1 scores between our algorithmically generated event segments and the ground truth segments (col 7 and col 12). F1 scores range between 0 and 1, and higher F1 scores imply greater overlap between our results and the ground truth segments. For the fine-grain action steps, we also report the average number of steps which have been undersegmented (col 8) and the average number of steps which have been oversegmented (col 9) when compared to the ground truth. We calculate undersegmentation as the number of algorithmically generated steps that overlap with multiple ground truth steps. We calculate oversegmentation as the number of ground truth steps that overlap with multiple algorithmically generated steps.

Pipeline	**Avg F1**	**Range**
(1) Video Shot Detn	0.37	0.3 - 0.44
(2) Text Phrase Detn	0.24	0.15 - 0.37
(3) Shot Detn + Phrase Detn	0.54	0.34 - 0.75
(4) Shot Detn + Phrase Detn + Video Shot Lbl	0.61	0.37 - 0.80
(5) Shot Detn + Phrase Detn + Text Phrase Lbl	0.77	0.67 - 0.87
(6) Full Pipeline	0.79	0.64 - 0.90

Table 3: Ablating components of Phase 1 of our fine-grain segmentation algorithm. All of the ablated pipelines include the Construct Shot-phrase Pairs component of Phase 1 as well as the components listed in the table. Each ablated pipeline reduces the average $F1$ overlap scores (and the range of these scores) with respect to ground truth segmentations and the fine-grain action step level.

7.1 Study I: Comparison to YouTube Interface

7.1.1 Methods. For Study I (a 60-minute remote session), we recruited eight participants from our organization to follow two instructional makeup videos, one using our mixed media tutorial interface (Ours) and the other using the YouTube interface (YouTube) as a baseline condition. All participants were US based women, ages 20 to 35, who had experience using makeup and reported watching makeup tutorials at least once a week. We selected two videos of similar length (about 4 minutes) from Table 2 (examples B and C in Figure 9). To minimize effects of the variations in tutorial difficulty and style on user performance, we chose videos that shared the same creator and focused on similar makeup look (a smokey eye). In the condition using our interface, we provided participants with the instructions that were fully automatically generated: Tutorial B contained one case of undersegmentation and one case of oversegmentation; tutorial C contains two cases of undersegmentation and two cases of oversegmentation.

We counterbalanced the order of the videos and interfaces across participants. Participants had up to fifteen minutes to complete each task, immediately followed by a series of 5-point Likert-scale questions about their experience. At the end of the session, we provided a cumulative questionnaire to collect their feedback. To conduct the study sessions over video call, we shipped the same makeup tools and products to each participant ahead of time.

7.1.2 Findings. Overall, participants followed the instructions in both interfaces by watching one step at a time all the way through, rewinding to review the tools and products, and replaying the step to execute the action. Some participants watched all or part of a step a few times to refine and touch-up their initial application. Below we describe the detailed findings of participants' navigation strategies and the confidence levels.

Navigation. All participants found our interface easier to navigate than the YouTube baseline (Median=5, σ=0.46 for Ours vs. M=3.5, σ=0.74 for YouTube, and p=0.01 using a Wilcoxon test). Six of the eight participants felt that they were able to follow the tutorial at their own pace better using our interface. The remaining two participants felt that they maintained the same pace with both interfaces. Our step-by-step breakdown enabled participants to quickly replay or scrub within a step, which helped them concentrate more on each individual step. In comparison, participants found that the tutorial went by too quickly in the YouTube interface, forcing them to expend a lot of effort pausing and scrubbing backwards longer distances and more often than when using our interface. These navigation difficulties discouraged them from spending too much time on each step. All participants found it easier to replay steps (p=0.02) in Our UI (M=5, σ=0.35) than YouTube (M=3, σ=1.28).

Participants found jumping between steps easier (p=0.02) in our tool (M=5, σ=0.46) then using YouTube (M=3.5, σ=1.41). The

coarse-level segmentation by part-of-face groups helped participants narrow down their search space, and the action step summaries enabled them to quickly find the desired step within a group. Participants strongly agreed that the hierarchical presentation was helpful (M=5). Additionally, in instances where the creator referred to a product or color that she mentioned earlier, participants relied on the step summaries to identify the step to jump back to.

Participants appreciated having the clickable, time aligned transcript above the video as an alternative navigation method to scrubbing the video timeline and found it particularly useful for locating products. The clickable transcript also made it easier for participants to handle undersegmentation in our tool. Once they realized that they were watching two steps in one, they used the transcript to identify the boundary between the two steps and to navigate back to the start of a particular step. Participants were also easily able to handle oversegmentations in our tool. Once they realized that a step was incomplete, they simply clicked on the next step to continue the instruction. Participants felt that the automatic step segmentations were very accurate (M=5). Because these tutorials contain very few instances of missegmentation, participants were able to handle the errors using our interface. However, we acknowledge that a follow-up study is needed to investigate how users review a tutorial that contains more segmentation errors.

Confidence. Seven of the eight participants felt more confident about their performance (p=0.04) using Our UI (M=3.5, σ=1.31) than using YouTube (M=3, σ=1.07) , but they gave low confidence scores for both interfaces because they doubted their own makeup skills. The remaining participant said that following the first tutorial using our tool built up her confidence for completing the second task in the YouTube interface.

The overview timeline visualizes the coarse tutorial structure by showing how much of the tutorial is focused on each part of the face. Participants said it helped them gauge which face parts were most important for replicating the look. Participants found that the step-by-step breakdown made them more confident that they attempted all steps. Participants often read the transcript to ensure that they didn't miss important narration. In contrast, with the YouTube baseline, participants could not be sure if they missed a step or some important detail because they scrubbed over it or because it automatically played while they weren't paying attention. Participants also cited the ability to focus on individual steps in detail and to easily return to previous steps for refinements or touch-ups as another reason for the higher level of confidence with out interface. While they might have wanted to back-reference or touch up previous steps in the YouTube condition as well, participants shared that they often didn't because the navigation overhead was "too much work." Participants were also more confident that they were using the right makeup tool for a given step (p=0.09) in Our UI (M=4.5, σ=1.13) compared to the baseline (M=3.5, σ=1.3).

Overall, every participant preferred our interface to the YouTube baseline for following the makeup tutorials.

7.2 Study II: Understanding Hierarchy Benefits

7.2.1 Methods. In Study II, we invited the participants from Study I to join a 20-minute follow-up remote interview a week after Study I. Our goal was to understand the effects of the hierarchical structure of our tutorials on comprehension and navigation. Six of the eight participants completed Study II. Each participant was shown two versions of our interface: one containing the two-level hierarchical structure which they saw in Study I and a version that removed the coarse-level of the hierarchy (i.e., the grouping by facial parts and the overview timeline were missing). Each participant had five minutes to explore and interact with each interface. We then asked them to compare the two interfaces by answering the following questions, "Which version do you prefer and why?" and "How do you think your process would change using this version compared to the previous version?".

7.2.2 Findings. In Study II, all participants expressed preference for the two-level, hierarchically structured version of the interface over the one-level version showing only the fine-grain action steps. All participants appreciated that the hierarchically segmented version helped them to quickly identify which "facial parts the tutorial focused on" and how many action steps would be required to apply the makeup to those facial parts (e.g., how complicated it would be). Multiple participants shared that when they watch makeup tutorials, they don't always follow the entire tutorial, but are only looking to execute part of it, such as the specific eye or contour techniques). When reflecting back their experiences in Study I, they explained that our two-level hierarchy allowed them to find the sections of interest to them far more quickly than the baseline YouTube interface. Some participants preferred to break the sequential ordering of the original makeup video when following the tutorial. For example, one user preferred to apply her face makeup before her eye makeup, even though the tutorial showed the eye makeup application first. Our hierarchical UI let her easily identify the face makeup section, so that she could complete that section before going back to the earlier section on the eyes.

All participants found the hierarchy useful in helping them to gauge their progress through the video. Participants often used the part-of-face section headers in the steps panel to determine when they finished with one face part so that they could pause and jump backwards to previous steps to touch up this face part before moving onto the next. One participant shared that the coarse-level part-of-face groupings helped her identify a "good stopping point" to touch up her eye-shadow because she "could easily tell when [the creator] wasn't going to apply any more eye makeup" and she "wanted make sure the eyes looked good before moving onto the lips or face".

Multiple participants shared that the part-of-face groupings in the steps panel helped them better remember the instructions after the task. One participant said that being able to "picture the right panel with the headers for the face parts and then the individual snapshots with some of the text of what was happening in every step" enabled her to describe multiple steps of the tutorial in detail.

7.3 Other Usability Feedback

While none of our participants gave negative feedback related to the hierarchical structure, they did give lower-level suggestions to could increase usability. Many participants toggled back and forth between the beginning and end of a step to visualize the difference resulting from the makeup application for that step. To make this visualization easier, they suggested adding a comparison before and

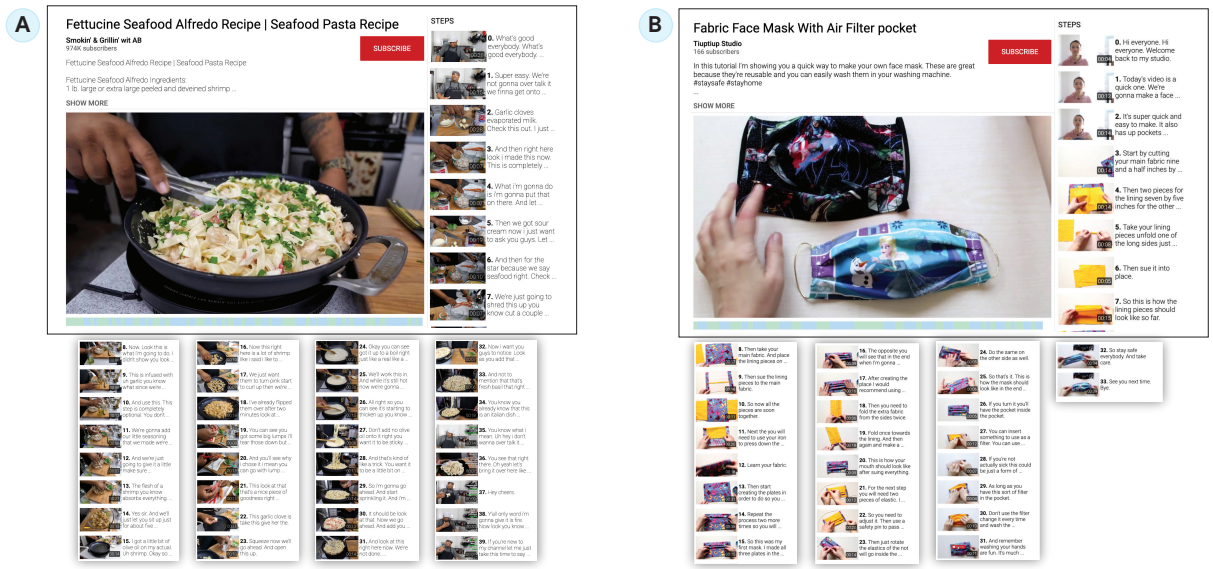


Figure 10: Two non-makeup mixed-media tutorials automatically generated by our pipeline. Tutorial A is an instructional cooking video, and tutorial B is a DIY video for sewing face masks. Since we do not compute coarse level groupings for non-makeup videos, only the finer-level action steps are visualized. ***Video Sources: (a) Fettucine Seafood Alfredo Recipe by Smokin' & Grillin' wit AB, (b) How to: FABRIC FACE MASK - Sewing Tutorial - ft. Filter Pocket by Craft with Laura. All licensed under CC BY.***

after image for each step. Additionally, when using the YouTube interface, participants found the preview frames displayed over the scrub bar to be extremely helpful and recommended that we add this feature to our video player as well. Finally, participants had varying comments on the UI layout, which suggests that we should enable customization of the layout for different users and contexts. For example, some participants wanted to see only the playback panel while following a step, but for the rest of the UI to be available by trigger, on pauses or while jumping between steps. Other participants liked to see all the panels for context even while following a step. One participant wanted to change the layout based on her comfort level with the tutorial: she preferred to hide the summary text for familiar makeup styles, but wanted the text as guidance for more complex tutorials.

8 LIMITATION AND FUTURE WORK

Results on other domains. While the focus of our work is on instructional makeup videos – in part because computer vision techniques are especially robust for detecting facial parts – we apply our algorithm to six non-makeup instructional videos and suggest how our system can be useful for general domains. We select these videos from the DIY, cooking, and bartending domains. We process these videos by removing any makeup-based heuristics, such facial part identification and product keywords, from our algorithm. Our algorithm cannot automatically perform the coarse level segmentation for these non-makeup tutorials. Figure 10 shows two of examples that we segmented: a cooking tutorial (Figure 10a) and a DIY face mask tutorial (Figure 10b). While the action steps (example a, steps 1-33 and example b, steps 3-25) for the instructional portion of the tutorial are sensible, we heavily over-segment the commentary section that follows (example a, steps 34 - 40 and example b, steps 25-33). Despite this over-segmentation and lack of hierarchy, we suggest that our automatic results are still valuable to creators as a first pass approximation for the fine-grain action steps. An open challenge for these other domains is to identify the appropriate coarse level structure and develop automated techniques to extract this structure. We believe this coarse structure is domain specific and will require domain specific computer vision techniques for automatic extraction.

Generalizability of face-based algorithmic components. In addition to the makeup tutorials shown in this paper, our techniques can also be applied to other types of facial makeup application such as theater or Halloween makeup. However, the face tracking technology may fail when makeup looks are too extreme (i.e., drawing additional eyes onto the face). The face-based components of our algorithm can also be applied to other tutorials domains such as facial hair grooming, facial expression acting, whistling, facial rehabilitation, and speech articulation.

Supporting multi-person videos. In this work, we focus on videos where the creators apply makeup on themselves. While this is the most common setup amongst YouTube makeup tutorials, another setup exists where a makeup artist applies the makeup to a model that often involve more conversations and interaction [24]. Currently, our approach does not support these multi-person setups as it would require more scene understanding of the people in the frame and their roles. However, it is feasible to extend our approach in future work.

Personalizing tutorial content. In our user study, some of the participants had facial features that were very different from the creator's. These participants expressed that when they normally watch makeup tutorials, they often have to experiment with how to adjust the look for themselves. We imagine extending our work

so that, in addition to breaking down the steps, our pipeline also synthesizes how each step and the final look on the user. Active research on makeup transfer and AR makeup applications would assist in this effort [4, 6, 22].

9 CONCLUSION

Makeup videos are one of the most prevalent forms of instructional videos on the Web today. We have demonstrated that these videos are often organized using a two-level object-action hierarchy with coarse-grained events focusing on facial parts and fine-level actions focusing on applying makeup products to parts of the face. We show that we can extract this hierarchy using multi-modal analysis of the video frames and text analysis of the transcript and that visualizing the hierarchy as mixed-media tutorial facilitates following the instructions. We believe that extracting and visualizing the object-action hierarchy can be applied to many other domains to make it easier for people to learn and follow different types of procedural tasks.

ACKNOWLEDGMENTS

We thank Tiffany Lee, Karen Kavett and our user study participants for their valuable insights. This work has been possible thanks to the support of people including, but not limited to the following (in alphabetical order of last name): Yuan Hao, Yury Kartynnik, Austin Myers, Bo Pang, Justin Parra, Gokul Raghuraman, Mogan Shieh, Chen Sun, Kanstantsin Sokal, and Andrey Vakunov. The first author is supported by the Brown Institute for Media Innovation.

REFERENCES

[1] Maneesh Agrawala, Doantam Phan, Julie Heiser, John Haymaker, Jeff Klingner, Pat Hanrahan, and Barbara Tversky. 2003. Designing Effective Step-by-step Assembly Instructions. *ACM Trans. Graph.* 22, 3 (July 2003), 828–837. https://doi.org/10.1145/882262.882352

[2] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. 2016. Unsupervised learning from narrated instruction videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE, Las Vegas, NV, USA, 4575–4583.

[3] B. Brown. 2008. *Bobbi Brown Makeup Manual: For Everyone from Beginner to Pro.* Grand Central Publishing, New York, NY, USA.

[4] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. 2018. PairedCycleGAN: Asymmetric Style Transfer for Applying and Removing Makeup. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE, Salt Lake City, UT, USA, 40–48.

[5] Minsuk Chang, Anh Truong, Oliver Wang, Maneesh Agrawala, and Juho Kim. 2019. How to Design Voice Based Navigation for How-To Videos. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, Article 701, 11 pages. https://doi.org/10.1145/3290605.3300931

[6] Hung-Jen Chen, Ka-Ming Hui, Szu-Yu Wang, Li-Wu Tsao, Hong-Han Shuai, and Wen-Huang Cheng. 2019. BeautyGlow: On-Demand Makeup Transfer Framework With Reversible Generative Network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE, Long Beach, CA, USA, 10042–10050.

[7] Pei-Yu Chi, Sally Ahn, Amanda Ren, Mira Dontcheva, Wilmot Li, and Björn Hartmann. 2012. MixT: Automatic Generation of Step-by-Step Mixed Media Tutorials. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (Cambridge, Massachusetts, USA) *(UIST '12)*. Association for Computing Machinery, New York, NY, USA, 93–102. https://doi.org/10.1145/2380116.2380130

[8] Pei-Yu Chi, Joyce Liu, Jason Linder, Mira Dontcheva, Wilmot Li, and Bjoern Hartmann. 2013. DemoCut: Generating Concise Instructional Videos for Physical Demonstrations. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology* (St. Andrews, Scotland, United Kingdom) *(UIST '13)*. Association for Computing Machinery, New York, NY, USA, 141—-150. https://doi.org/10.1145/2501988.2502052

[9] Jonathan D. Denning, William B. Kerr, and Fabio Pellacini. 2011. MeshFlow: Interactive Visualization of Mesh Construction Sequences. In *ACM SIGGRAPH 2011 Papers* (Vancouver, British Columbia, Canada) *(SIGGRAPH '11)*. Association for Computing Machinery, New York, NY, USA, Article 66, 8 pages. https://doi.org/10.1145/1964921.1964961

[10] Logan Fiorella and Richard E Mayer. 2018. What works and doesn't work with instructional video.

[11] C. Ailie Fraser, Joy O. Kim, Hijung Valentina Shin, Joel Brandt, and Mira Dontcheva. 2020. Temporal Segmentation of Creative Live Streams. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376437

[12] Google. 2020. *Cloud Natural Language documentation | Cloud Natural Language API.* Google Cloud. https://cloud.google.com/natural-language/docs

[13] Google. 2020. *Cloud Speech-to-Text - Speech Recognition | Google Cloud.* Google Cloud. https://cloud.google.com/speech-to-text

[14] Floraine Grabler, Maneesh Agrawala, Wilmot Li, Mira Dontcheva, and Takeo Igarashi. 2009. Generating Photo Manipulation Tutorials by Demonstration. In *ACM SIGGRAPH 2009 Papers* (New Orleans, Louisiana) *(SIGGRAPH '09)*. Association for Computing Machinery, New York, NY, USA, Article 66, 9 pages. https://doi.org/10.1145/1576246.1531372

[15] Tovi Grossman, Justin Matejka, and George Fitzmaurice. 2010. Chronicle: Capture, Exploration, and Playback of Document Workflow Histories. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology* (New York, New York, USA) *(UIST '10)*. Association for Computing Machinery, New York, NY, USA, 143–152. https://doi.org/10.1145/1866029.1866054

[16] Julie Heiser, Doantam Phan, Maneesh Agrawala, Barbara Tversky, and Pat Hanrahan. 2004. Identification and Validation of Cognitive Design Principles for Automated Generation of Assembly Instructions. In *Proceedings of the Working Conference on Advanced Visual Interfaces* (Gallipoli, Italy) *(AVI '04)*. ACM, New York, NY, USA, 311–319. https://doi.org/10.1145/989863.989917

[17] D. Huang, J. J. Lim, L. Fei-Fei, and J. C. Niebles. 2017. Unsupervised Visual-Linguistic Reference Resolution in Instructional Videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation / IEEE, Honolulu, HI, USA, 1032–1041.

[18] De-An Huang, Shyamal Buch, Lucio Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. 2018. Finding "It": Weakly-Supervised Reference-Aware Visual Grounding in Instructional Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE, Salt Lake City, UT, USA, 5948–5957.

[19] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. 2017. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. Computer Vision Foundation / IEEE, Honolulu, HI, USA, 7310–7311.

[20] R. Jones. 2017. *Robert Jones' Makeup Masterclass: A Complete Course in Makeup for All Levels, Beginner to Advanced.* Fair Winds Press, Beverly, MA, USA.

[21] Yoshinari Kameda and Michihiko Minoh. 1996. A human motion estimation method using 3-successive video frames. In *International conference on virtual systems and multimedia*. VSMM, Gifu, Japan, 135–140.

[22] Archana Kannan. 2020. Shopping for a beauty product? Try it on with Google. https://blog.google/products/shopping/shopping-beauty-product-try-it-google/

[23] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J. Guo, Robert C. Miller, and Krzysztof Z. Gajos. 2014. Crowdsourcing Step-by-Step Information Extraction to Enhance Existing How-to Videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 4017–4026. https://doi.org/10.1145/2556288.2556986

[24] Bridjet Lee and Kasia Muldner. 2020. Instructional Video Design: Investigating the Impact of Monologue- and Dialogue-Style Presentations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376845

[25] Rainer W. Lienhart. 1998. Comparison of automatic shot boundary detection algorithms. In *Storage and Retrieval for Image and Video Databases VII*, Minerva M. Yeung, Boon-Lock Yeo, and Charles A. Bouman (Eds.), Vol. 3656. International Society for Optics and Photonics, SPIE, San Jose, CA, USA, 290 – 301. https://doi.org/10.1117/12.333848

[26] Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nick Johnston, Andrew Rabinovich, and Kevin Murphy. 2015. What's Cookin'? Interpreting Cooking Videos using Text, Speech and Vision. *CoRR* abs/1503.01558 (2015), 1–10. arXiv:1503.01558 http://arxiv.org/abs/1503.01558

[27] Julie Bauer Morrison, Barbara Tversky, and Mireille Betrancourt. 2000. Animation: Does it facilitate learning. In *AAAI spring symposium on smart graphics*, Vol. 5359. The AAAI Press, Menlo Park, CA, USA, 8.

[28] Megha Nawhal, Jacqueline B Lang, Greg Mori, and Parmit K Chilana. 2019. VideoWhiz: Non-Linear Interactive Overviews for Recipe Videos.. In *Graphics*

Interface. Canadian Information Processing Society, Kingston, Ontario, Canada, 15–1.

[29] Celie O'Neil-Hart. 2017. Self-directed learning from YouTube - Think with Google. https://www.thinkwithgoogle.com/advertising-channels/video/self-directed-learning-youtube/

[30] 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* (Honolulu, Hawaii, USA) *(UIST '14)*. ACM, New York, NY, USA, 573–582. https://doi.org/10.1145/2642918.2647400

[31] Promise Phan. 2015. 'INSIDE OUT' Makeup Tutorial (Disgust,Sadness,Joy,Anger & Fear). https://www.youtube.com/watch?v=C1DXqkOCBt0

[32] Suporn Pongnumkul, Mira Dontcheva, Wilmot Li, Jue Wang, Lubomir Bourdev, Shai Avidan, and Michael F. Cohen. 2011. Pause-and-Play: Automatically Linking Screencast Video Tutorials with Applications. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) *(UIST '11)*. Association for Computing Machinery, New York, NY, USA, 135–144. https://doi.org/10.1145/2047196.2047213

[33] Ozan Sener, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. 2015. Unsupervised semantic parsing of video collections. In *Proceedings of the IEEE International Conference on Computer Vision*. Computer Vision Foundation / IEEE, Santiago, Chile, 4480–4488.

[34] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*. Computer Vision Foundation / IEEE, Long Beach, CA, USA, 7464–7473.

[35] Anh Truong, Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. 2016. QuickCut: An Interactive Tool for Editing Narrated Video. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 497–507. https://doi.org/10.1145/2984511.2984569

[36] Sylvaine Tuncer, Barry Brown, and Oskar Lindwall. 2020. On Pause: How Online Instructional Videos Are Used to Achieve Practical Tasks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376759

[37] Sylvaine Tuncer, Barry Brown, and Oskar Lindwall. 2020. On Pause: How Online Instructional Videos Are Used to Achieve Practical Tasks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376759

[38] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE, Columbus, OH, USA, 1386–1393.

[39] Sarah Weir, Juho Kim, Krzysztof Z. Gajos, and Robert C. Miller. 2015. Learnersourcing Subgoal Labels for How-to Videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 405–416. https://doi.org/10.1145/2675133.2675219

[40] Wikipedia contributors. 2020. F1 score. https://en.m.wikipedia.org/wiki/F1_score

[41] Ann Yuan and Andrey Vakunov. 2020. Face and hand tracking in the browser with MediaPipe and TensorFlow.js. https://blog.tensorflow.org/2020/03/face-and-hand-tracking-in-browser-with-mediapipe-and-tensorflowjs.html

[42] Jeffrey M Zacks and Barbara Tversky. 2001. Event structure in perception and conception. *Psychological bulletin* 127, 1 (2001), 3.

[43] Jeffrey M Zacks, Barbara Tversky, and Gowri Iyer. 2001. Perceiving, remembering, and communicating structure in events. *Journal of experimental psychology: General* 130, 1 (2001), 29.

[44] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. 2019. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Computer Vision Foundation / IEEE, Long Beach, CA, USA, 3537–3545.

Title (Duration)	Creator	URL	Date	Referencing Figures
(A) Yellow Cut Crease Makeup (807s)	PatrickStarrr	https://youtu.be/9f7zmCSzG9E	May 16, 2016	Tbl 3, Figs 6, 9
(B) Smokey Eye Makeup Look (279s)	TheMakeupChair	https://youtu.be/pOMc7gp0KHg	Nov 8, 2019	Tbl 3, Figs 5, 6, 8, 9
(C) Glam Makeup (234s)	TheMakeupChair	https://youtu.be/zbNfJRxvkV8	Jun 28, 2019	Tbl 3, Figs 2, 9
(D) Morphe Brushes 35R Palette (423s)	Brittney Enora	https://youtu.be/Fh6olbNwFbI	Apr 23, 2017	Tbl 3, Figs 5, 9
(E) 1920s Wearable Makeup (358s)	Jbunzie	https://youtu.be/CMtd-rJcE1k	Jun 11, 2018	Tbl 3, Figs 5, 9
(F) Nude Makeup (272s)	corallista	https://youtu.be/GUayJLkqVY8	Aug 5, 2015	Tbl 3, Fig 9
(G) Day to Night Makeup (492s)	corallista	https://youtu.be/yYOfKGls2Dc	Aug 27, 2015	Tbl 3
(H) How to Apply Flawless Foundation (778s)	Brittney Enora	https://youtu.be/3UhMh3dT2ME	Apr 13, 2018	Tbl 3, Fig 6
(I) Teal Glitter Smokey Eye (666s)	ParticKStarrr	https://youtu.be/pDe_hDjUC74	Dec 23, 2015	Tbl 3, Fig 1
(J) Easy Glowy Summer Makeup (321s)	BeKami	https://youtu.be/SsyLv2hCZpw	Apr 20, 2018	Tbl 3, Fig 5
(K) Fettucine Seafood Alfredo	Smokin' & Grillin' wit AB	https://youtu.be/BgGmWwXW69A	Dec 9, 2020	Fig 10
(L) Fabric Face Mask ft. Filter Pocket	Craft with Laura	https://youtu.be/OUYJlKo5qtE	Mar 22, 2020	Fig 10

Table 4: Video Credits. We would like to thank all the creators listed whose videos we show in this paper. All videos here are licensed with CCBY Creative Commons.