Multimodal Analysis: Researching Short-form Videos and the Theatrical Practices[1]

YITING WANG[2] [https://orcid/org/0000-0002-5407-0085]
University of Hawai'i at Mānoa

Citation
Wang, Y.. (2021). *Multimodal Analysis: Researching Short-form Videos and the Theatrical Practices*. Proceedings of the 104th Association for Education in Journalism and Mass Communication. http://hdl.handle.net/10125/76003

Abstract
Video analysis methods need to be updated in facing the proliferation of user-generated short-form videos (UGSVs). This paper investigates what's behind this type of visual communication and resonates with the emerging field of social media and performance studies. We ask if a theatrical or performative discourse can make sense of the video data, and search for a method to holistically study UGSVs.

This paper uses multimodal analysis for video analysis and draws concepts and practices from Chinese and Western theater. Building on three theories (situation, suspense, and mimesis) in which the ontology of theater is often discussed, this paper demonstrates the modes and modalities of five videos originating from TikTok. The preliminary findings suggest three types of suspense and three types of mimesis practices that respectively answer how attention of audiences is retained, and how and why videos are reproduced and disseminated. We argue that imitation as a phenomenon and as a process can generate memes, and memes in turn invites more imitation. The underlying crux are the video practices that ridicule and critique, when different levels of resistance to politics, authority, or societal classes are shown.

Video analysis, under today's ubiquitous visual data, requires robust updates. In addition to the contribution of a performative and theatrical perspective for the sense-making of short-form videos, this paper also contributes to the methods of video analysis in general and video analysis by using modes and multimodalities.

Keywords
Multimodal Analysis, short video analysis, theater, performance, TikTok

---

[1] Top student paper award, visual communication division.
[2] Yiting Wang: yitingw@hawaii.edu, Interdisciplinary Ph.D. Program in Communication and Information Sciences

Introduction

Short-form video is one of the major approaches with which people present themselves on visual social media, such as TikTok. Methods to study these videos; however, are very much needed and in their early development. We are in search of a method that goes beyond videography, that can investigate video content on the one hand, while on the other provoking the hidden theory.

Multimodal analysis comes to handy for both purposes in that multimodal communication focuses on the medium of the expression first and the communicative practice second. For this paper, rather than treating video as videos, we parse video as a sequence of moving images (Lienhart et al., 1997, p. 55). This reduces the complications of analysis, handles rich video data well, and is certainly more manageable when analyzing short-form videos.

Looking at these user-generated short-form videos (UGSVs), we have discovered modes and modalities as evidence that connect with a theatrical system. We found modes being repetitively used throughout the video data, and their connections with three theatrical theories in which the ontology of theater is often discussed (situation, suspense and mimesis). By demonstrating five UGSV examples, we concluded that there are three types of suspense that accompany situations and they explain audience engagement; three types of mimesis that link with the mimetic and meme-like repetition and representations; and their connections with compositional, social and political modalities. Short-form videos practices are personal, yet political. They show different levels of resistance and emancipation via humor to ridicule, and to respond to social phenomena.

The three theatrical theories for analyzing UGSVs are new in the literature but a performative perspective in social media is not. As Schechner (2020) states, digital living is performative (p. 283); as Lonergan (2016) suggests, social media is a performance space; and as Yarosh et al. (2016) discovered, 62% of the youth-authored short-form videos are "staged, scripted or choreographed". With the increasing interest of social media in the performance studies, this paper contributes to the literature by using a theatrical perspective to analyze visual data. What's more, it contributes not only to multimodal analysis of UGSVs, but also to the video analysis in general methodologically.

Short-form videos

Short-form videos as a medium, stresses the shortness in length; such expression has sprung up over the years. Commonly found in Chinese literature, yet inconsistent in English and Chinese context, the earliest definitions focus on videos that take seconds as the length indicator, usually user generated, and rely on mobile intelligent terminals for quick shooting, editing and sharing. The length limit is debatable between 15 to 10 minutes (Interactive

Advertising Bureau [IAB], 2009; SocialBeta, 2015), and the latest publications suggest 15 to 60 seconds, and not longer than 5 minutes (Kong, 2018; Chen et al., 2020). Sometimes referred to as "short video" (Zhou et al., 2018), or "instant video clip" (Zhang et al., 2014); however, microcontent, or micro-media more holistically describes such phenomenon. Microcontent, a term first introduced in 1998, now describes "information published in short form…, constrained by software and devices that we use to view digital content". Digital microcontent at the same time is a media phenomenon, producing new forms of media content, new practices, new experiences, and new forms of circulations (Lindner & Bruck, 2007, p. 56).

Theater

Performance studies have increasingly become interested in its relationship with social media. This paper focuses on theater studies, an offshoot of performance studies for several reasons.

First, performance remains broad, but theater is specific. As Carlson (2017) defines: "all activities carried out with a consciousness of itself" can be performance. Conversely, theater shows clear modes: "text, mise-en-scène, lighting, casting, music, effects, placing on the stage, the nature of the audience, the price and availability of tickets, pubs or cafes, and the relationships between all these considerations" (McGrath, 1981, p. 5). Second, performance does not necessarily specify the existence of an audience. Third, theater emphasizes the sense of space where a performance is reenacted at.

Hence, theater refers to "the complex of phenomena associated with the performer-audience transaction, with the production and communication of meaning in the performance itself and with the systems underlying it" (Elam, 2002, p. 2). Here, the system refers to the whole system of coordination of performers, directors, technicians and so forth. Drama, to clarity, is "the mode of fiction designed for the purpose of stage presentation" (Elam, 1980, p. 2). A reader may see theater, performance and drama used in the writing interactively; they each carry different intentions as defined above.

Situation & Suspense and Mimesis (S&S-M)

What do we talk about when we talk about theater? Theater theorists have three theories regarding the ontology of theater: situation & suspense (S&S); mimesis and audience (Tan, 2005). For this paper, the focuses are the first two: S&S and mimesis.

Situation & Suspense (S&S)

During the global COVID-19 pandemic where face covering is mandated, a TikToker narrates the experience in a video: "I was grocery shopping, and I was wearing a mask. I felt something was wrong on my face, but I couldn't take off my mask [due to COVID risks]. I

kept shopping but my face was getting tingly. Then my lips went numb. I quickly finished shopping, tried hard not to rip off my mask and rushed to the car. I closed the door, threw my stuff away, and took my mask off. *There was a spider in my mask*".

Though a straight-forward narration, this video perfectly demonstrates two core elements in drama: situation & suspense. The narrator builds up a set of situations (i.e., feeling wrong, tingly face, numb lips) to develop a pressing storyline, to reveal the suspense (spider) which retains the audience's attention. This is the situation & suspense. Situation is the essence of theater or drama, while suspense is naturally generated in the growth of situations. The two are indivisible. Situations trigger the audience to explore urgently about the next possible action for a situated performer, provoking suspense.

Mimesis

Among UGSVs, Sarah Cooper's impersonation of Trump by using his voice and improvising with body gestures to match the content of his speech, is an example of imitation. This paper employs mimesis, which is the reference of imitation in theater. The difference is that mimesis not only addresses imitation as repetition, but also representation (Rasmussen, 2008). Defined as "a multi-leveled series of repetitions and reproductions" (Gruber, 1987, p. 204), mimesis associates with human phenomena such as the "acts of mimicry, imitative social behavior, the performance and re-performance of identity, and the summoning of otherness in the medium of the self" (Larlham, 2012).

In fact, the origin of traditional Chinese theater connects with imitation closely. During the period of late Western Zhou (1047 BC - 772 BC), an occupation "you" appeared for sensual pleasures in imperial court. "You" imitates to remonstrate tactfully, or to mock unreasonable behaviors of rulers without concerning penalization (Niu, 2004, p. 8). The premise of "you" is that by gaining pleasure out of watching the imitation, people contemplate, learn, or infer (Aristotle, 1922, p. 15).

Multimodal Analysis

Short-form videos are full of modes. What are modes? In short-form videos, modes are every perceivable little element. As introduced by Kress (2009), modes are semiotic resources, they refer to a medium of communication such as speech or writing (Dicks, 2019, p. 3); mode can also be "resource for making meaning", they are linguistic or nonlinguistic, written or spoken, acoustic or visual. Multimodal analysis tries to capture and analyze how participants use specific resources for meaning-making: hand gestures, bodily movements, objects that were brought in the participation frame (Dicks, 2019).

The concept of "multimodality" dates back to the 1920s (Leeuwen, 2020) and was developed in the early 2000s (Jewitt, 2016, p. 70). Multimodality is defined as "the use of several semiotics modes in the design of a semiotic product or event". To communication

practices, a multimodal approach is to examine textual, aural, linguistic, spatial, and visual resources (Pearce et al., 2018, p. 6). Jewitt (2016, p. 69) states that digital technologies are of particular interest to multimodality because they make available a wide range of modes, often in new inter-semiotic relationships with one another. They unsettle and re-make genres, and they often reshape practices and interaction.

A multimodal theory of communication focuses on two things: 1) the semiotic resources of communication, the modes and the media used; 2) the communicative practices in which these resources are used (Kress & Van Leeuwen, 2001, p. 111). This aligns closely with the five steps this paper employs for multimodal analysis: sampling data, transcribing data, analyzing individual modes, analyzing across modes and connecting them with theories. The first four links with the semiotic resources of communication; the last step explains the communicative practices.

For simplicity, we will display five videos in total, all originating from TikTok platform (two from the TikTok website, one featured on the news, and two from our own data collection in 2019). The purpose is to demonstrate the method and invite future replications.

First, we shall investigate the five steps:

1. Sampling data. Videos are dynamic and malleable, thus creating challenges for analysis; but videos are a set of images, which allows researchers to break down videos as a sequence of images. That is the agenda for sampling. One can sample according to time, or the start/end of an event, or by thematic ideas or theoretical concepts (Kress et al., 2005). TikTok videos were limited to 15 seconds in length at first, then extended to 60 seconds. For this paper, the chosen time interval is 15 seconds. The criterion for sampling is the emergence of storylines, instructed by S&S-M. The five videos cited as examples are sampled manually. That is, we break down the frames by hand.

2. Transcribing data. This step is to capture all the salient modes and semiotic resources in sequential order, then translate the video to texts into a storyboard. It may involve "noting the use of different modes, semiotic resources, and modal affordances" (Jewitt, 2016, p. 78). For instance, how the objects are placed, or how people gestured in the video. In this step, we do not analyze modes yet; instead, this is a mode discovery phase. The ultimate goal is to map out all the major and salient modes and the most important job for this step is to introduce the storyline in the video. In that without the video, one can still roughly know the major events. This will be fully illustrated in the examples. One shall not expect to exhaust all the modes; rather, it is an iterative process in steps.

3. Analyzing individual modes. Unlike merely displaying mode in step 2, this step is to analyze individual modes and raise questions such as: what work is being done by each of the modes? When and why is one mode used over another? What is the primary mode in this

video frame? What affordances are being exploited and for what social purpose? During this step, we will also enrich our coding dictionary, based on step 2.

4. Analyzing across modes. This step is to build a multimodal ensemble which shows the interactions and representations constructed by one or more modes (Jewitt, 2016). Moving from individual mode to multimodal ensemble is essential because it is the multimodal layering that showcases the social use of modes, in that modes have different affordances and materialities (Jewitt, 2016, p. 82). This step asks researchers to analyze across modes altogether and raise questions such as: in this video, åhow do the combination of texts, facial expressions, props, and spatial relationships work together?

5. Connect with theory. Multimodality deeply cares about how the communication and interactions between modes connect to a systematic explanation. That is why the last step is to combine multimodality with (social) theories to explain communicative practices. For this step, we ask, what kind of theory can we employ to make sense of the phenomena constructed by these modes? It is extremely important to note that this step can be employed as early as any inquiry occurred in any aforementioned steps, as explained by Jewitt (2016, p. 83). For this paper, in this step, we see modes connecting with S&S-M, which tackles our research question: how does a theatrical or performative system help to make sense of short-form videos?

Multimodal Analysis of Five Videos

Now we have understood the exact five steps. In this section, we will analyze five videos to demonstrate multimodal analysis.

Video analysis 1 - *Diner* [3]

Recorded from TikTok website, named *Diner* (Figure 1), the first video is 12 seconds long. The video ends with 11 frames for analysis (sampling) and each frame provides various modes. Table 1 provides detailed textual descriptions (Storyline, data transcribing), which made up for the limitation of a paper that is restricted in texts. In this table, F means frame and each number corresponds to each frame. Storyline corresponds to step 2, data transcribing; and Individual Modes displays step 3, analyzing individual mode. The analyses are conducted in Excel. The mode abbreviations will be used throughout the paper.

Figure 1. ***Diner***

[3] TikTok [@TikTok]. *Our Mission*. [Video]. TikTok. https://www.tiktok.com/about?lang=en. Partial synopses see Figure 2.

Table 1. Video 1 Analysis (*Diner*)

F	Storyboard (Data Transcribing)	Individual Modes
1	Diner smells food, using hands to waft the smell.	Body movement (BM)
2	Interaction: eye contact, okay hand sign	Hand gesture (HG), facial expression (FE), spatial relationship (SR)
3	Fling open paper napkin w/ one hand	Prop (P)
4	Tuck napkin as bib, dining	P, BM
5	Show food (fries, nuggets, burgers cut in halves, jam); Show silverware, cut food w/ silverware	P
6	Dressed up waiter (bow tie, shirt, towel on lower arm; jeans and canvas shoes), consults for drink	SR, P
7	Diner shows approval	BM
8	Sommelier pour (can of soda) in wine glass	BM, HG, P
9	Sniff, sip, pinky out	HG
10	Tilt head, shrug, curl lips	FE, BM
11	Silverware away, shovel food in mouth w/ hands	BM

Putting all the mode data together, in this video, a performer sits at the table with a fine bib (an unfolded napkin) and fine silverware to eat burgers, French fries and nuggets. A semi-well-dressed waiter appears with a towel hanging on the lower arm and shows a can of

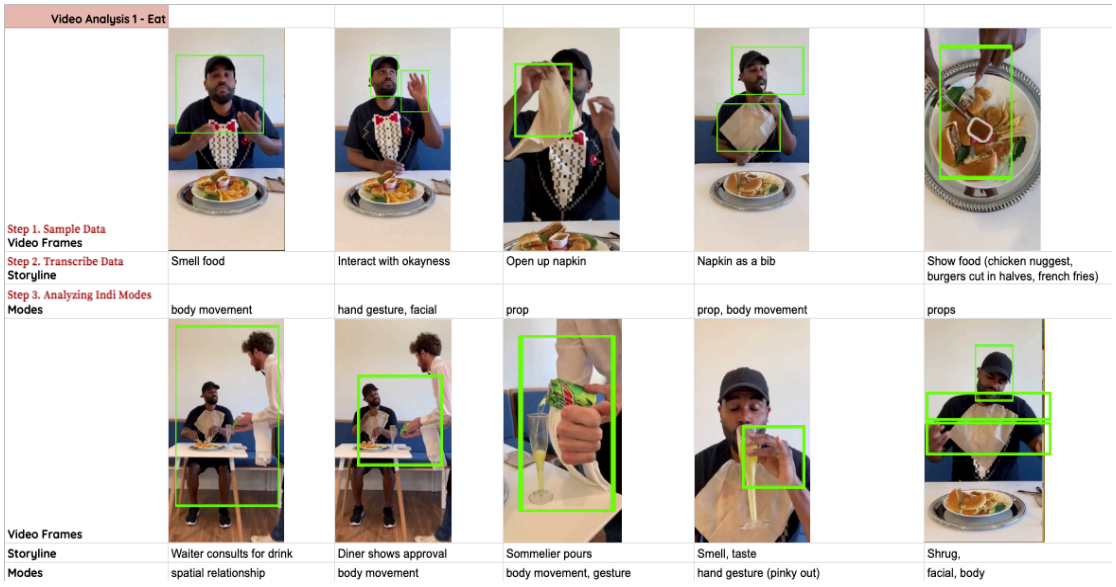

Figure 2. Multimodal Analysis for Video *Diner*

This figure demos the first three steps for multimodal analysis.

soda in the way of showing fine wine. The diner signals approval and satisfaction, the waiter

maneuvers a professional sommelier pour. The diner appreciates the drink, then shovels the food in the mouth with his hands.

In this video, there are many different modes added in the mode dictionary: body movement (BM), hand gesture (HG), facial expression (FE), props (P), and spatial relationship (SR). How do all the modes function together and how do they connect with theories?

In *Diner*, Modes in frame 1 to 9 help to *build situations*: smell with enjoyable facial expressions (BM, FE); okay sign to show gratification, eye direction shows another performer exists outside of the frame (HG, SR, FE); use silverware, napkin (bib), for burgers, fries and nuggets (P); waiter comes in, consult drink, show approval, sommelier pour, pinky out (SR, BM, HG, P, FE). Frame 10 and 11 *transit to suspense*: tilted head, shrug, curl lips, then silverware away, shovel food in mouth with hands. The previous nice preparation, the serious interaction between diner and waiter, was broken by the final decision of shoveling food. This suspense resonates with previously set up situations: cutting fries and burgers with silverware, pouring soda in a wine glass.

Modes SR, BM, HG, P and FE resonate with imitation too. The diner imitates what a fine table manner would be like, pinky out for aristocracy; the waiter imitates sommelier pour, props such as bow tie and towels to make-believe a professional waiter in an expensive restaurant. Imitations here functioned as make-believe but not make-belief. Audiences are aware of the boundary between a theatrical performance and everyday reality (Schechner & Brady, 2013, p. 42).

Video analysis 2 - *Air Dancer* [4]

Recorded from TikTok website, named *Air Dancer* (Figure 3), this 8 second video is sampled into 6 frames. This video is a pure imitation of air-dancers by using body movements and color specified props and costumes. The situation is very linear (two people imitate air-dancers) without a clear suspense. Therefore, salient modes for this video, in addition to body movements and props, are costume (COS) and color (CLR). The costumes two performers wear are color specified to suit the color of two air-dancers (props): one performer in magenta to suit the magenta long big balloon with human faces; another in orange to suit the orange air-dancer. Table 2 shows more details.

Figure 3. ***Air Dancer***

[4] TikTok [@TikTok]. Our Mission. [Video]. TikTok. https://www.tiktok.com/about?lang=en.

Table 2. Video 2 Analysis (*Air Dancer*)

F	Storyboard (Data Transcribing)	Individual Modes
1	1st performer in orange twists body, waves arms	BM, Costume (COS, orange t-shirt, shorts)
2	1st Air dancer appears, twists body	Color (CLR, orange), P
3	2nd performer in magenta repeats F1	BM, COS (magenta t-shirt, shorts)
4	2nd air dancer appears, twists body	CLR (magenta), P
5	2 Performers wave/dance together	F1, F3
6	2 air-dancers wave/dance together	F2, F4

Video analysis 3 - *Me/Her Dad* [5]

Retrieved from our 2019 data collection, named *Me/Her Dad* (Figure 4), this 10 second video is sampled into 5 frames. Previously used modes are BM, SR, and CLR, and new modes are text (T), text box (TB), background music (BGM) and mise-en-scène (MSC).

Table 3. Video 2 Analysis (*Me/Her Dad*)

F	Storyboard (Data Transcribing)	Individual Modes
1	Performer ties shoelace next to ocean, distracts by waves	HG, text (T), text box (TB), CLR, mise-en-scène (MSC)
2	Ocean waves come, performer dodges in hurry	SR, BGM ("f*** u"), T, TB
3	Waves chase, text box shows the simile: ocean waves show how "her dad" feels about "me"	SR, BM, T, TB, BGM ("I hate your friends")
4	Performer totters	Same as F3
5	Performer flees helter-skelter	BM, BGM ("and they hate me too")

In this video, the situations are built by the mise-en-scène (ocean and wave), the texts and text boxes strengthen the situation (F1-2) by locating the "Me" box next to the performer, and "Her Dad" against the ocean which creates a simile. The suspense comes when the wave and background music come in (F2-5) to reveal a story between a daughter's father and her partner. This simile uses the sudden wave to imitate an unwelcoming father towards his daughter's partner. The BGM may hint to imitate a father's internal mental activity (Table 3).

Figure 4. *Me/Her Dad*

[5] TikTok [@zaydelie]. (2019, September 22). *When you're at her house and her dad comes home* [Video]. TikTok. https://www.tiktok.com/@zaydelie/video/6739458795463855365?lang=en&is_copy_url=1&is_from_webapp=v1 For more information, see the last section *Limitations and Future Research*.

Video analysis 4 - *Africa in 1400* [6]

Reported by the *Time* magazine (Greenspan, 2019), named *African in 1400* (Figure 5), this 13 second video is sampled into 9 frames.

Table 4. Multimodal Analysis Step 1-3 (Africa in 1400)

F	Storyboard (Data Transcribing)	Individual Modes
1	Africa in 1400, Africans enjoying their daily hunting, about to shoot spear	T, TB, FE (concentrated), P (spear), COS (traditional), MSC (bedroom), ambient BMG
2	Focus, shush	HG (fingers to lips)
3	Heard something at the back	BGM ("wassp")
4	Started turning head	BGM ("oh lord, jetson made another one")
5	Turn and see, spear lowered	SR, FE (bothered, confused), P (spear), active beats BGM, BM.
6	Same performer now British people, intruded in Africa	P (sunglasses), T (British), TB/CLR (red), BM (hip-hop dance), COS (modern, crewneck sweatshirt, pants w/ Adidas logo), BGM (drop, "walkin").
7	Same performer now French, intruded in Africa	P (sunglasses), T (French), TB/CLR (purple), BM (hip-hop dance), COS (modern, pull over hoodie, hood on), BGM (drop, "walkin").
8	Same performer now Spaniards, intruded in Africa	P (sunglasses, baseball cap), T (Spaniards), TB/CLR (yellow), BM (hip-hop dance), COS (modern, zip-up hoodie, pants w/ Adidas 3 stripes), BGM (drop, "walkin").
9	Same performer now Portuguese, intruded in Africa	P (sunglasses), T (Portuguese), TB/CLR (pine green), BM (hip-hop dance), COS (modern, tank-top, pants w/ Adidas 3 stripes), BGM (drop, "walkin").

Figure 5. ***Africa in 1400***

In this video, modes help to build major situations and major suspense, accompanied by minor situations and minor suspense. F1-2 (Table 4) build a basic situation where Africans enjoy their daily hunt, F3-5 foreshadow the suspense when the performer clearly was distracted by something: a new situation in which British people intruded. French people's intrusion breaks it into a new situation, then two more countries - Spain and Portugal. New intrusion keeps changing the situations when a new country comes in and brings new suspense. This is an example of how situation and suspense (F6-9) always accompany each other as stated before.

[6] TikTok [@sharoonbi]. (2019, October 3). *Things weren't the same after that* [Video]. TikTok. https://www.tiktok.com/@sharoonbi/video/6743685493684260102?lang=en&is_copy_url=0&is_from_webapp=v1&sender_device=pc&sender_web_id=6891456073962997254

Some modes function as indispensable elements for building meaningful situations. The major situations are built by the 6 texts in their color-coded text boxes (F1-9): Africa in 1400s, Africans enjoying their daily hunt, British, French, Spaniards, and Portuguese. The absence of them will create a more difficult meaning-making. The second important modes for not only building but meaning-making the situation are the body movements, facial expressions, props, and costumes. Facial expression in F5 is pivotal for a suspense building in that the situation in transit is signaled well. The alternation between costumes and props, on the one hand, traditional hunting tool (spear) and indigenous cloth; on the other hand, hoodie, sunglasses, tank-top, Adidas pants, mark one situation after another. These modes also imitate the historical indigenous dressing and hunting. The imitation of hip-hop dance and mannerism may intend to display the detested arrogance and the infallibility of being colonizers. Lastly, the body movement also marks situations: moving the spear back and forth to estimate the hunting, in comparison to the hip-hop dance.

Figure 6. *Kids Asleep/Awake*

Video analysis 5 - *Kids Asleep/Awake* [7]

Retrieved from our 2019 data collection, named *Kids Asleep/Awake* (Figure 6), this 14 second video is sampled into 5 frames. Modes we have seen are body movement, spatial relationship, text, text box, and background music; the new mode is sound effect (SE).

Table 5. Multimodal Analysis Step 1-3 (*Kids Asleep/Awake*)

F	Storyboard (Data Transcribing)	Individual Modes
1	Performer tiptoed in the room	BM, BGM ("Twinkle little star")
2	Performer shushes (finger to lips)	HG, T ("when kids are sleeping")
3	Performer walks, opens door, lower body w/utmost care	BM, T, HG, sound effect (SE, door creaks)
4	Performer stomps, jumps, waves, cheers, laughs	BM, T ("when my husband and I are sleeping and my kids are up"), BGM (pop music)
5	Performer kicks balls, balloons, jumps, waves	BM, BGM (pop), SE (glass shatter, muffled balloon kick)

[7] TikTok. [@unknown]. (Yeah, Month Date unknown). Title unknown. [Video]. TikTok.

In Table 5, F1 to F3 introduces the situation where kids are asleep, F4 to F5 illustrates another situation where kids are awake, but parents are sleeping. The BGM contrasts with each other, along with the sound effect. The two modes imitate careful parents and reckless children in everyday life. The second textual description (kids asleep parents up), is a situation *and* a suspense. Figure 7 below gives a full display of video analysis in Excel.

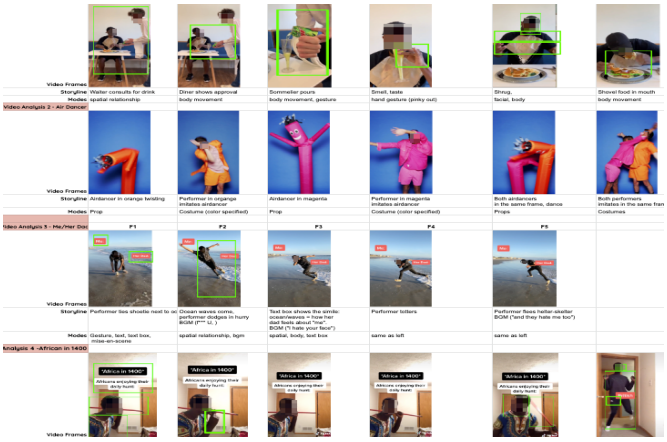

Figure 7. Multimodal Analysis in Excel

Preliminary Findings

Looking at the multimodalities in these short-form videos, we notice that on the face value, combining visual communication and performance studies, modes in digital performances and modes in traditional theatrical performances resonate; meanwhile, multimodalities and fundamental theatrical theories can work together to make sense of short-form videos. We also see the same modes being repetitively yet interchangeably used for situation building, suspense transition, or for imitation. Scaling up each mode and across modes, the results yield three types of suspense accompanying situations, and three types of imitation.

Three types of suspense accompanying situations

Using text and/or text boxes mode(s) to create S&S is salient among UGSVs. Sometimes users apply different colors for T or TB. This type of suspense that accompanies situations are *texto-visual suspense.* As analyzed before, the absence of text (or text boxes) will create equivocality. In *Me/Her Dad*, the situation where the performer tilts the body to avoid the ocean wave is a mere situation; what really illustrates the major expression in the video are the textual modes "Me" and "Her Dad", highlighted in red, which contrasts with the mise-en-scène mode that is blue. T and TB modes are the textual suspense, that pushes a situation further: a simile describing "my relationship with my partner's father". Otherwise, without such suspense, the video is a situation where a performer tries to avoid ocean waves.

T and TB are not exclusively only for suspense building; they can create situations. In video 5 *Kids Asleep/Awake*, the textual descriptions primarily set up a situation: when kids are asleep, which contrasts with another situation: when kids are awake, but parents are asleep. Strictly speaking, the second situation is also a suspense, because it reveals the turning-point of the story.

Aside from T and TB, the color of T and TB assists texto-visual suspense building as well. In *Africa*, the alteration between TB in red, purple, and yellow shows new situations one after another. Modes are not monomodal, they function differently in accordance with their composition. One mode serves one role in a frame may serve a different role in another frame.

Texto-audio suspense, is powered by mode BGM (sound and lyrics) and SE (e.g., door creaks). In *Africa in 1400*, when each country enters, altering the situations successively, the repetitive beats became suspense. Without the BGM mode, the video can still be going, but the whole premise of building-up a peaceful hunting situation to the continuous turned-over impede situation may be lost. The texto-audio mode (i.e., lyrics, "wassp", "walkin", "jetson made another one") within BGM functions as the overlapping sound (OS) of the situation when intruders came in one by one, in a casually "walked in" manner. The texto-audio suspense plays a pivotal role in the transitions of situations. Similarly, in *Me/Her Dad*, the texto-audio modes ("f*** u,", "and they hate me too") assist as suspense to transit the situation.

Device as suspense, mainly powered by modes prop (P) and costume (COS). As a theater term, prop is short for property. In *Africa in 1400,* physical props include P and COS choices such as indigenous cloth versus modern cloth, spear versus sunglasses, or BM, hip-hop dance versus traditional hunting movements. Digital devices may refer to video filters, or camera language. Similar to how suspense functions in drama, device suspense controls the major thesis of a video. A single filter can transfer a banal video to a theatrical digital performance. For example, Chinese users on Douyin (Chinese version of TikTok) apply head decoration filters to show their conduction of a traditional Chinese theater story. Device referring to artificial objects stems from Gibson (2015, p. 37) who coined affordances. We fully discuss the theatrical-level affordances and multimodalities in another paper and will not elaborate here (Wang & Suthers, 2021); however, we adopt Leonardi's (2010) proposal about digital materiality that "whether in physical or digital form, an artifact that translates idea into action is material" (p. 15).

Three types of mimesis

Audio, visual and devices modes also assist mimesis. Some commonly seen modes for mimesis are usually body movements (BM), hand gestures (HG), background music (BGM), or sound effect (SE). BM is commonly used for *mannerism* imitation. In the *Air Dancer* video, two performers twist their bodies to copy the movements to simulate manners of air dancers. In *Diner*, BM, HG, P resonate with the imitation of table manners, or experience in fancy restaurants. In *Kids,* the mannerism includes tiptoeing, kicking toys which relates to parents and children activities.

The second type include sound and/or sound effect. They relate to audio and is twofold: sound imitation (the sound is missing) and sound matching (the sound is given, commonly known as lip-syncing). When the sound is missing, performers imitate sound via onomatopoeia or the tone of speaking. For example, during festivals, when people can't celebrate with real fireworks due to the ban, a surge of UGSVs are posted online (e.g., Douyin) where performers stood tall and straight to imitate different types of fireworks. This resonates with *kouji* (imitate sounds with mouth) in folk vocal art form in traditional Chinese theater arts.

When the sound is given, in the imitative activities, performers use BM to accompany the exact mouthing of lyrics of popular songs, speeches, or dialogues created by them or other performers. The aforementioned Sarah Copper matching Trump's speech is an example.

The third type is *imitation with devices*: hand gestures, props, costumes and color as modes facilitate imitation. When pretending to be the air-dancer, or pretending to be children, creators refrain from make-belief; instead, they make-believe. Differentiated by Richard Schechner, a scholar in performance studies, make-belief is when a performer trying hard to convince an audience of the authenticity, ideally achieving a complete trust, adoration or belief. An advertisement of a product is to make-belief to win costumers. Conversely, make-believe is prone to be pretentious, audiences usually know the boundary between the reality and the drama. A staged performance usually is to make-believe. In UGSVs, a make-believe example can be that when picking props pretending on the phone, performers hold phone cases, TV remotes, or even shoes.

Discussion

At this moment, we understand that modes individually and collectively are able to set up situations and foreshadow suspense. Modes in short-form videos are capable of building theatrical relationships. In detail, texto-visual, texto-audio, and devices modes build three types of suspense that accompanies situation and three types of imitations. However, a pressing question is, how do these findings connect with theories?

Rose (2007) proposes four sites (approaches) in her visual methodology, and the object (image, video) itself is our major focus. In each site, Rose encourages scrutinizing different modalities: social, technological, or compositional. Modality here means three different aspects to a critical understanding of images (p. 25). It refers to the representation but not the authenticity of a visual object (Ledin & Machin, 2020, p. 63).

Situation and suspense are the compositional modalities, the two *compose* and explain how short-form-video-dramatists engage audiences. "Dramatists have no time to waste", as Baker (1919, p. 16) states. This is true for traditional stage and digital stage (short-form videos) as media. TikTok videos create situations at all times. Within these situations, the critical

suspense happens. Thus within 15-60 seconds, an audience is struck to not swipe away. Usually, due to time limitations for situation growth, short-form videos only have one suspense. Users accomplish one suspense within one or more situation(s). Their ultimate goal is to engage or "win the attention of the audience as promptly as possible" (Baker, 1919).

Mimesis connects more with social and political modalities. As mentioned earlier, mimesis can be pure repetition, the repetitive dance movements in TikTok videos can be proof. However, one shall not narrow or impoverish mimesis, as Halliwell (2002, p. 14) suggests. Beyond reproducing actions, mimesis scales up as representation. Representation calls for audiences to decipher the intended meaning with the video expressions, in that communication is meaning-embedded, concealed, or abstract (Rasmussen, 2008, p. 313). For example, the sound matching of Trump, mimicking fireworks, imitating the dining manner of elitism, making a phone call with a shoe, in different extent, showcase different levels of resistance to politics, societal class gradations, and the pressure of socialization. These videos too, show resistance to authority and carry the subtle sense of emancipation (Li et al., 2019).

In the sense of cultural or biological modalities, mimesis has a meme-like function. Memes, explained by Dawkins (2006, p. 192), are about ideas that leap from one brain to another via a process called imitation. Shifman (2013) also agrees that memes are produced by various means of imitation. We argue that imitation powers up memes and the nature of memes fuels more imitation. A meme replicates by infection (e.g., humor) of human minds to induce people to repeat its pattern (He & He, 2003). This contributes to explain the recurrent "challenges" phenomenon on TikTok that are disseminated, replicated, reproduced by users one after another.

Through multimodal analysis, we illustrate a theatrical framework that helps to make sense of short-form videos: situation & suspense (S&S) and mimesis. This system consists of various modes and multimodalities including compositional, social, and political. Along the way, we also discovered three types of suspense used in the video with different situations, and three types of mimesis. Situation and suspense, compositionally, answers how and why short-form videos engage audiences; mimesis, on the other hand, reflects on the short-form video practices politically, culturally, and socio-economically. Mimesis as imitation, as representation, and as memes, explains the recurrent interactions between performers: imitation connects and assists the growth of memetic culture. Memes are generated out of imitation as a process, and the nature of memes invites more imitation.

Limitations and Future Research

From a method perspective, the challenge of UGSVs is the abundance of modes. The dissemination across platforms, and data scraping may lead to the disappearance of mode. This may impact data analysis. For example, in some UGSVs (e.g., *Me/Her Dad*), the original video has an additional textual description outside of the video that was not included in the data scraping process. We treated each video with the same given information to analyze regardless (i.e., to not look into the additional text). It is possible to double check modes from the original resource but for a big corpus of data, it is challenging.

In data analysis, manual sampling also has the same problem when handling large dataset, which is why automatic sampling should be used. Though cannot elaborate due to space limitation, we did sample automatically using a python script (Gordon, 2020). Automatic sampling also creates challenges such as imprecise time interval or missed important modes. Future research can apply the theatrical framework in automatic sampling, which proves to be helpful via manual sampling.

This method-centric paper aims to introduce ways to analyze UGSVs, thus theory elaborations are limited (e.g., discussions about memes). Nevertheless, questions such as the materiality of a mode, the affordances (e.g., mimeticality) that lay in the materiality should be asked**.** More specifically, how to define materiality in a digital environment? Are the props or costumes used in the videos, digital or physical (Rathnayake & Suthers, 2018)? What are the potentials and limitations of a mode (Kress, 2010, p. 84)? Such questions lead to discussions about a performative or theatrical relationship between users and the platform which can be referred to as theatrical-level affordance. This also speaks to the long-standing relationship between affordance and multimodality.

Bibliography

Aristotle. (1922). *The Poetics of Aristotle* (S. H. Butcher, Trans.; 4th ed.). Macmillan and Co., Limited.

Baker, G. (1919). *Dramatic Technique*. The Riverside Press, Cambridge.

Carlson, M. (2017). *Performance: A critical introduction* (Third edition). Routledge.

Chen, X., Kaye, D. B. V., & Zeng, J. (2020). # *PositiveEnergy* Douyin: Constructing "playful patriotism" in a Chinese short-video application. *Chinese Journal of Communication*, 1–21. https://doi.org/10.1080/17544750.2020.1761848

Dawkins, R. (2006). *The selfish gene* (30th anniversary ed). Oxford University Press.

Dicks, B. (2019). *Multimodal Analysis.pdf*. SAGE Publications Ltd.

Elam, K. (1980). *The semiotics of theatre and drama*. Methuen.

Elam, K. (2002). *The semiotics of theatre and drama* (2nd ed). Routledge.

Gibson, J. J. (2015). *The ecological approach to visual perception*. Psychology Press.

Gordon, M. (2020). *Automatic Video Frame Capture* [Python].

Greenspan, R. (2019). *Teens Are Getting Millions of Views With Theatrical History Lessons on TikTok. These Real Historians Are Thrilled.* Time. https://time.com/5721116/teen-tik-tok-history-lessons-videos/

Gruber, W. E. (1987). "Non-Aristotelian" Theater: Brecht's and Plato's Theories of Artistic Imitation. *Comparative Drama*, *21*(3), 199–213. https://doi.org/10.1353/cdr.1987.0007

Hadley, B. (2017). *Theatre, social media, and meaning making*. Springer Berlin Heidelberg.

Halliwell, S. (2002). *The aesthetics of mimesis: Ancient texts and modern problems*. Princeton University Press.

He, Z., & He, X. (2003). Memetics and Social Usage of Language. *Modern Foreign Languages (Quaterly)*, *Vol. 26*(No. 2).

Interactive Advertising Bureau. (September, 2009). *Long Form Video Overview*. IAB Interactive Advertising Bureau.

Jewitt, C. (2016). Multimodal Analysis. In A. Georgakopoulou & T. Spilioti (Eds.), *The Routledge handbook of language and digital communication*. Routledge.

Knoblauch, H., Tuma, R., Atkinson, P., Delamont, S., Cernat, A., Sakshaug, J. W., & Williams, R. A. (2020). *Videography and video analysis*. https://methods.sagepub.com/foundations/videography-and-video-analysis

Kong, D. (2018). *Research Report on Short Video Industry*. 36kr Research Center. http://www.199it.com/archives/672181.html

Kress et al., G. R. (2005). *English in urban classrooms: A multimodal perspective on teaching and learning*. RoutledgeFalmer.

Kress, G. (2009). What is Mode? In C. Jewitt (Ed.), *The Routledge handbook of multimodal analysis*. Routledge.

Kress, G. R. (2010). *Multimodality: A social semiotic approach to contemporary communication*. Routledge.

Kress, G. R., & Van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. Arnold ; Oxford University Press.

Larlham, D. (2012). *The Meaning in Mimesis: Philosophy, Aesthetics, Acting Theory* [Columbia University]. https://academiccommons.columbia.edu/doi/10.7916/D8B27SBG

Li, M., Tan, C. K. K., & Yang, Y. (2019). *Shehui Ren*: Cultural production and rural youths' use of the *Kuaishou* video-sharing app in Eastern China. *Information, Communication & Society*, 1–16. https://doi.org/10.1080/1369118X.2019.1585469

Larlham, D. (2012). The Meaning in Mimesis: Philosophy, Aesthetics, Acting Theory [Columbia University]. https://academiccommons.columbia.edu/doi/10.7916/D8B27SBG

Ledin, P., & Machin, D. (2020). *Introduction to multimodal analysis*. Bloomsbury Academic.

Leeuwen, T. van. (2020). *The SAGE Handbook of Visual Research Methods* (L. Pauwels & D. Mannay, Eds.). SAGE Publications, Inc. https://doi.org/10.4135/9781526417015

Leonardi, P. M. (2010). Digital materiality? How artifacts without matter, matter. *First Monday*, *15*(6). https://doi.org/10.5210/fm.v15i6.3036

Li, M., Tan, C. K. K., & Yang, Y. (2019). *Shehui Ren*: Cultural production and rural youths' use of the *Kuaishou* video-sharing app in Eastern China. *Information, Communication & Society*, 1–16. https://doi.org/10.1080/1369118X.2019.1585469

Lienhart, R., Pfeiffer, S., & Effelsberg, W. (1997). Video abstracting. *Communications of the ACM*, *40*(12), 54–62. https://doi.org/10.1145/265563.265572

Light, B., Burgess, J., & Duguay, S. (2018). The walkthrough method: An approach to the study of apps. *New Media & Society*, *20*(3), 881–900. https://doi.org/10.1177/1461444816675438

Lindner, M., & Bruck, P. (2007). *Micromedia and Corporate Learning*. Proceedings of the 3rd International Microlearning 2007 Conference.

Lonergan, P. (2016). *Theatre and social media*. Palgrave Macmillan.

McGrath, J. (1981). *A good night out: Popular theatre: audience, class, and form*. Eyre Methuen.

Niu, B. (2004). *History of Chinese Traditional Opera*. Culture and Art Publishing House.

Pearce, W., Özkula, S. M., Greene, A. K., Teeling, L., Bansard, J. S., Omena, J. J., & Rabello, E. T. (2018). Visual cross-platform analysis: Digital methods to research social media images. *Information, Communication & Society*, *0*(0), 1–20. https://doi.org/10.1080/1369118X.2018.1486871

Rathnayake, C., & Suthers, D. D. (2018). Twitter Issue Response Hashtags as Affordances for Momentary Connectedness. *Social Media + Society*, 14.

Rasmussen, B. (2008). Beyond imitation and representation: Extended comprehension of mimesis in drama education. *Research in Drama Education: The Journal of Applied Theatre and Performance*, *13*(3), 307–319. https://doi.org/10.1080/13569780802410673

Rose, G. (2007). *Visual methodologies: An introduction to the interpretation of visual materials* (2nd ed). SAGE Publications.

Schechner, R. (2020). *Performance studies: An introduction* (Fourth edition). Routledge, Taylor & Francis Group.

Schechner, R., & Brady, S. (2013). *Performance studies: An introduction* (3rd ed). Routledge.

Shifman, L. (2013). Memes in a Digital World: Reconciling with a Conceptual Troublemaker. *Journal of Computer-Mediated Communication*, *18*(3), 362–377. https://doi.org/10.1111/jcc4.12013

SocialBeta. (2015). *Marketing Guidance for Short-Form Videos.* https://socialbeta.com/t/short-video-marketing-guide-2015.html

Tan, P. S. (2005). *Theater Ontology*. China Theater Publishing House. China Xiju Publishing House.

Wang, Y., & Suthers, D. (2021). *Understanding Affordances for Theatricality and Performativity in Short-Form Videos*. Communication and Information Sciences PhD Program, University of Hawai'i at Mānoa.

Yarosh, S., Bonsignore, E., McRoberts, S., & Peyton, T. (2016). YouthTube: Youth Video Authorship on YouTube and Vine. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16*, 1421–1435. https://doi.org/10.1145/2818048.2819961

Zhang, L., Wang, F., & Liu, J. (2014). Understand Instant Video Clip Sharing on Mobile Platforms: Twitter's Vine as a Case Study. *ACM Digital Library*, 85–90. https://doi.org/10.1145/2597176.2578278

Zhou, C., Zhong, S., Geng, Y., & Yu, B. (2018). A Statistical-based Rate Adaptation Approach for Short Video Service. *2018 IEEE Visual Communications and Image Processing (VCIP)*, 1–4. https://doi.org/10.1109/VCIP.2018.8698706