The Prose Storyboard Language: A Tool for Annotating and Directing Movies

Rémi Ronfard, Vineet Gandhi, Laurent Boiron, Vaishnavi Ameya Murukutla

Eurographics Workshop on Intelligent Cinematography and Editing (2022), pp. 1–9
R. Ronfard and H.-Y. Wu (Editors)

The Prose Storyboard Language: A Tool for Annotating and Directing Movies (Version 2.0, Revised and Illustrated Edition)

Rémi Ronfard[1] and Vineet Gandhi[2] and Laurent Boiron[3] and Vaishnavi Ameya Murukutla[1]

[1]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, France
[2]International Institute of Information Technology, Hyderabad, India
[3]Weta Digital, New Zeland

Abstract

The prose storyboard language is a formal language for describing movies shot by shot, where each shot is described with a unique sentence. The language uses a simple syntax and limited vocabulary borrowed from working practices in traditional movie-making and is intended to be readable both by machines and humans. The language has been designed over the last ten years to serve as a high-level user interface for intelligent cinematography and editing systems. In this new paper, we present the latest evolution of the language, and the results of an extensive annotation exercise showing the benefits of the language in the task of annotating the sophisticated cinematography and film editing of classic movies.

CCS Concepts

• ***Applied computing*** → ***Media arts;***

1. Introduction

In movie production, directors often use a semi-formal idiom of natural language to convey the shots they want to their cinematographer. Similarly, film scholars use a semi-formal idiom of natural language to describe the visual composition of shots in produced movies to their readers. In order to build intelligent and expressive virtual cinematography and editing systems, we believe the same kind of high-level descriptions need to be agreed upon. In this paper, we propose a formal language that can serve that role. Our primary goal in proposing this language is to build software cinematography agents that can take such formal descriptions as input, and produce fully realized shots as an output. A secondary goal is to perform in-depth studies of film style by automatically analyzing real movies and their scripts in terms of the proposed language.

The prose storyboard language is a formal language for describing shots visually. We leave the description of the soundtrack for future work. The prose storyboard language separately describes the spatial structure of individual movie frames (compositions) and their temporal structure (shots). In film analysis, there is frequent confusion between shots and compositions. A *medium shot* describes a composition, not a shot. If the actor moves towards the camera in the same shot, the composition will change to a *close shot* and so on. Therefore, a general language for describing shots cannot be limited to describing compositions such as *medium shot* or *close shot* but should also describe screen events which change the composition during the shot. In the prose storyboard language, each shot is a complete sentence with at least one composition and an arbitrary number of screen events. This offers unprecedented expressive power for describing and directing movies.

Our language can be used indifferently to describe shots in preproduction (when the movie only exists in the screen-writer and director's minds), during production (when the camera records a continuous "shot" between the times when the director calls "camera" and "cut"), in post-production (when shots are cut and assembled by the film editor) or to describe existing movies. The description of an entire movie is an ordered list of sentences, one per shot. Exceptionally, a movie with a single shot, such as Rope by Alfred Hitchcock, can be described with a single, long sentence.

In this paper, we assume that all shot descriptions are manually created. We leave for future work the important issue of automatically generating prose storyboards from existing movies, where a number of existing techniques can be used [Bra97, DMR05, GCSS06, GR13]. We also leave for future work the difficult problems of automatically generating movies from their prose storyboards, where existing techniques in virtual camera control can be used [HCS96, CAwH*96, JY06, CO06, GCR*13, GRCS14, GCLR15].

2. Prior art

Our language is loosely based on existing practices in moviemaking [TB09b, TB09a] and previous research in the history of

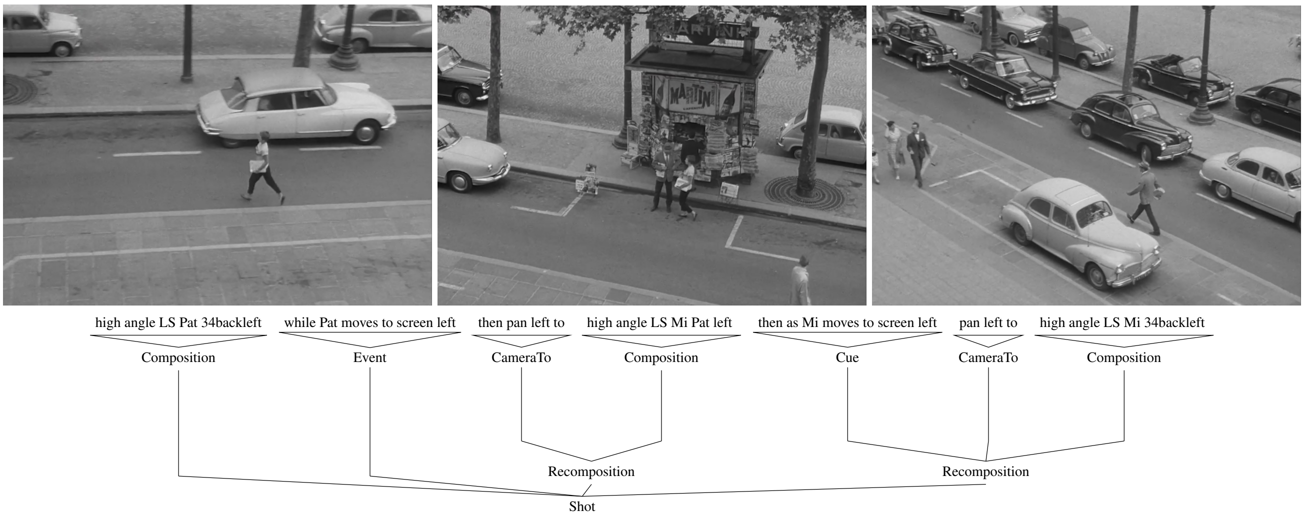

high angle LS Pat 34backleft
while Pat moves to screen left
then pan left to
high angle LS Mi Pat left
then as Mi moves to screen left
pan left to
high angle LS Mi 34backleft
Composition
Event
CameraTo
Composition
Cue
CameraTo
Composition
Recomposition
Recomposition
Shot

Figure 1: *Complex shot in* Breathless

film style [Bor98, Sal09]. Our language is also related to the common practice of graphic storyboards. In a graphic storyboard, each composition is illustrated with a single drawing. The blocking of the camera and actors can be depicted with a conventional system of arrows within each frame, or with a separate set of floor plan views, or with titles between frames.

In our case, the transitions between compositions use a small vocabulary of screen events including camera actions (pan, dolly, crane, hold, continue) and actor actions (speak, react, move, cross). Although the action vocabulary could easily be extended, we voluntarily keep it small because our focus in this paper is restricted to the blocking of actors and cameras, not the high-level semantics of the narrative.

We borrow the term prose storyboard from Proferes [Pro08] who used it as a technique for decomposing a films script into a sequence of shots, expressed in natural language. Other authors use the French term "decoupage" to describe this important step in film production. The name catches the intuition that the language should directly translate to images. In contrast to Proferes, our prose storyboard language is a formal language, with a well-defined syntax and semantics, suitable for future work in intelligent cinematography and editing.

Our proposal is complementary to the Movie Script Markup Language (MSML) [RKV*09], which encodes the structure of a movie script. In MSML, a script is decomposed into dialogue and action blocks but does not describe how each block is translated into shots. Our prose storyboard language can be used to describe the blocking of the shots in a movie in relation to an MSML-encoded movie script. For this purpose, MSML makes provision for a Manufacturing model and Animation model, which are only loosely described. The prose storyboard language can be seen as an alternative representation for both the Manufacturing and Animation models in MSML.

Our proposal is also related to the Declarative Camera Control Language (DCCL) which describes film idioms, not in terms of cameras in world coordinates but in terms of shots in screen coordinates [CAwH*96]. The DCCL is compiled into a film tree, which contains all the possible editings of the input actions, where actions are represented as subject-verb-object triples. Our prose storyboard language can be used in coordination with such constructs to guide a more extensive set of shot categories, including complex and developing shots.

Our approach is also related to the work of Jhala and Young who used the movie Rope by Alfred Hitchcock to demonstrate how the storyline and the director's goal should be represented to an automatic editing system [JY06]. They used Crossbow, a partial order causal link planner, to solve for the best editing, according to a variety of strategies, including maintaining tempo and depicting emotion. They demonstrated the capability of their solver to present the same sequence in different editing styles. But their approach does not attempt to describe the set of possible shots. Our prose storyboard language attempts to fill that gap.

Other previous work in virtual cinematography [SMAY03,JY05, FF06, ORN09, MJSB11, GRCS14, GCLR15, LDTA17] has been limited to simple shots with either a static camera or a single uniform camera movement. Our prose storyboard language is complementary to such previous work and can be used to define higher-level cinematic strategies, including arbitrarily complex combinations of camera and actor movements, for most existing virtual cinematography systems.

Text-to-movie (or text-to-scene) authoring is a general class of methods that have been proposed for automatically generating 3D graphics and animation from natural language text. Good results have been obtained in limited domains, such as generating 3D scenes from natural language accident reports [ASSN03] or generating cartoon animation from scripted dialogue scenes [SY06]. Commercially available text-to-movie systems include NawmalMAKE and Plotagon Studio Such systems use marked-up dialogues as input and generate simple shots with minimal camera movements and editing. Other related work along the same lines includes Ye and Baldwin [YB08] who describe a method for generating storyboards from movie scripts; Marti et al. [MVW*18] who

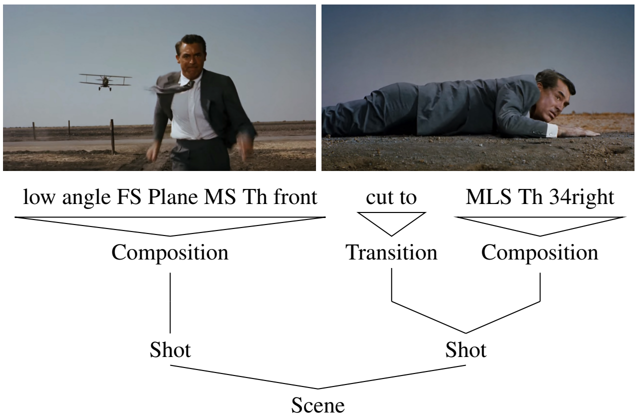


Figure 2: *Prose storyboard language description of two iconic shots in Alfred Hitchcock's* North By Northwest

describe methods for generating previz animation, also from movie scripts.

Closer to our approach, Director Notation [Yan13] is a symbolic language intended to express the content of film (motion pictures), much as musical notation provides a language for the writing of music. But DN is a graphical notation whereas PSL is a pseudo-natural language, and DN describes the movie production process, whereas PSL describes the movie itself, as in a storyboard. TIMISTO [VHL*13] and SLAP [BS16] are pattern languages for creating animation from storyboards.

A previous version of the prose storyboard language was presented at the international workshop on intelligent cinematography and editing (WICED) in 2012. Since then, several variations of PSL have been used for generating cinematic replays in serious games [GRCS14], for generating synthetic complex shots from live video material [GRG14, GR15], for staging complex scenes in 3D animation [LCL18], for directing cinematographic drones [GLC*18] and for learning film editing patterns from examples [WPRC18]. In this revised edition, we provide a definitive version of the language, illustrated with a large number of examples and a reference implementation, in the hope that it will stimulate future work in automatic annotation and generation of movies.

3. Requirements

The prose storyboard language is designed to be expressive, i.e. it should describe arbitrarily complex shots, while at the same time being compact and intuitive. Our approach has been to keep the description of simple shots as simple as possible, and allowing for more complex descriptions when needed. Thus, for example, we describe actors in composition from left to right, which is an economical and intuitive way of specifying relative actor positions in most cases. As a result, our prose storyboard language is very close to natural language (see Fig.21 for examples).

It should be easy to parse the language into a non-ambiguous semantic representation that can be matched to video content, either for the purpose of describing existing content or for generating novel content that matches the description. It should therefore be possible (at least in theory) to translate any sentence in the language into a sketch storyboard, then to a fully animated sequence.

It should also be theoretically possible to translate existing video content into a prose storyboard. This puts another requirement on the language, that it should be possible to describe existing shots just by watching them. There should be no need for contextual information, except for place and character names. As a result, the prose storyboard language can also be used as a tool for annotating complete movies and for logging shots before post-production. Since the annotator has no access to the details of the shooting plan, even during post-production [Mur86, Ond04], we must therefore make it possible to describe the shots in screen coordinates, without any reference to world coordinates.

4. Syntax and semantics

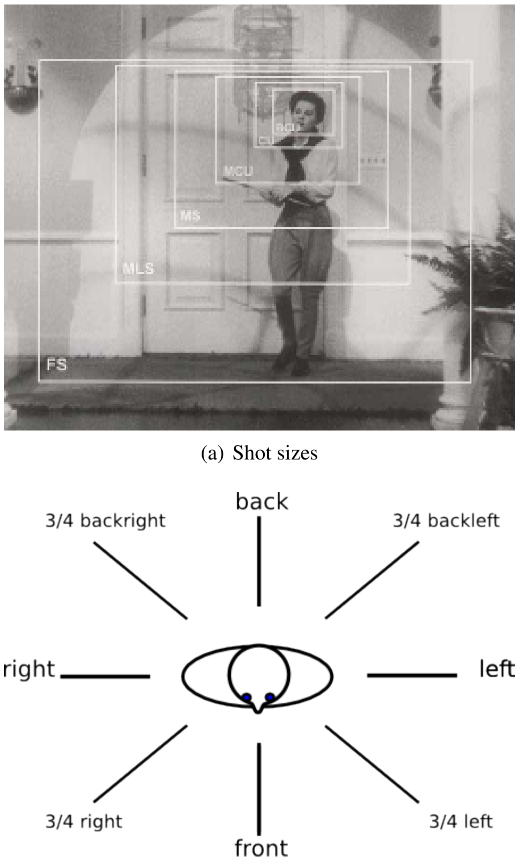
The prose storyboard language is a context-free language, whose terminals include generic and specific terms. Generic terminals are used to describe the main categories of screen events including camera actions (pan, dolly, cut, dissolve, etc.) and actor actions (enter, exit, cross, move, speak, react, etc.). Specific terminals are the names of characters, places and objects that compose the image and play a part in the story. Non-terminals of the language include important categories of shots (simple, complex, developing), image compositions and developments. The complete grammar for the language is illustrated with the AND/OR graph in Fig. 14 and described in the PEG notation in Fig.15. A reference implementation using the Parsimonious toolkit for parsing PSL sentences using the Python language can be found at https://gitlab.inria.fr/vmurukut/psl.

The semantics of the prose storyboard language is best described in terms of a Timed Petri Net (TPN) where durative events such as compositions and camera actions are represented as *places* ; and instantaneous events (such as cuts and the start and end of other events) are represented as *transitions*. We use timed Petri nets, rather than finite state machines, as a semantic representation of shots in a prose storyboard, in order to adequately represent developing shots with an elaborate choreography of actor and camera movements taking place simultaneously, such as the opening shot in Orson Welles' "Touch of evil".

TPNs have been proposed for representing the temporal structure of movie scripts [RKV*09], character animation [MRR98, BvKR01], game authoring [BBAG08], turn-taking in conversation [Cha12] and synchronisation and storage models for multimedia systems [LG90]. Our model differs from previous work by representing all durative events with places and using transitions to synchronize them. For lack of space, we differ the description of the Petri Net interpretation of the prose storyboard language to future work.

5. Image Composition

Image composition is the way to organise visual elements in the motion picture frame to deliver a specific message to the audience. In our work, we propose a formal way to describe image composition in terms of the actors and objects present on the screen and the spatial and temporal relations between them.



(a) Shot sizes

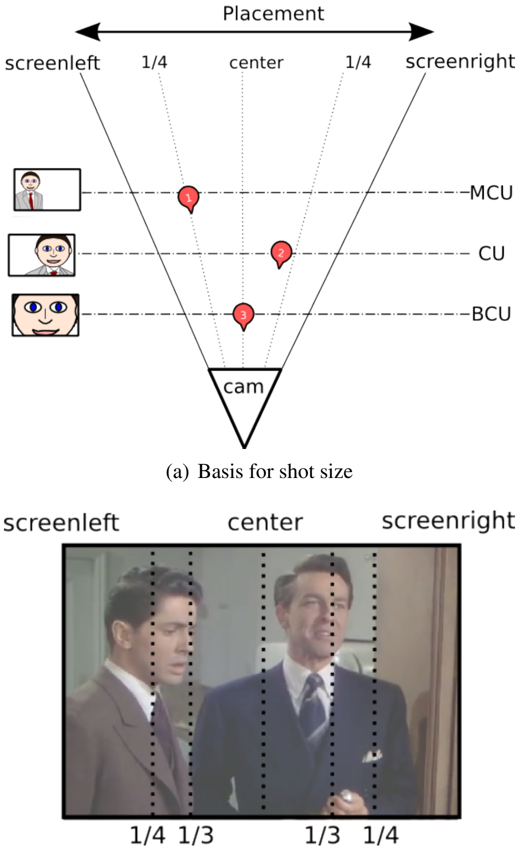
(b) Profile angle

Figure 3: *(a) shows shot sizes in the prose storyboard (reproduced from [Sal06]). (b) shows the profile angle of an actor defines his orientation relative to the camera. For example, an actor with a* left *profile angle is oriented with his left side facing the camera.*

Following Thomson and Bowen [TB09b], we define a composition as the relative position and orientation of visual elements called *Subject*, in screen coordinates. In the simple case of *flat staging*, all subjects are more or less in the same plane with the same size, but in the case of *deep staging*, different subjects are seen at different sizes, in different planes. For the sake of generality, we therefore choose to indicate the size of each subject separately. As a result, each subject is defined by its size, profile angle, and screen position. See Fig. 3(b) and 3(b) for a visual explanation.

As a convention, we assume that the subjects are described from left to right. This means that the left-to-right ordering of actors and objects is part of the composition. Indeed, because the left-to-right ordering of subjects is so important in cinematography and film editing, we introduce a special keyword *cross* for the screen event of one subject crossing over or under another subject. Shot sizes are used to describe the relative sizes of subjects independently of the camera lens as illustrated in Fig.4(a).

The *Screen* term describe the subject position in screen coordinate. It allows to slightly modify the generic framing by shifting the subject position to the left or right corner as shown in Fig.4(b).



(a) Basis for shot size

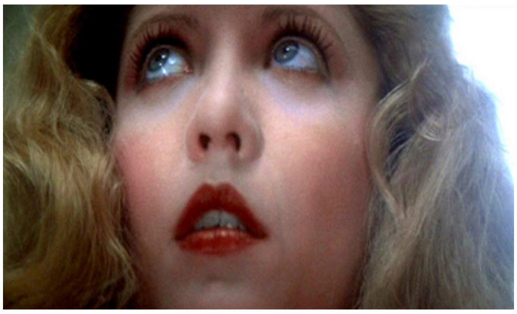
(b) Screen coordinates

Figure 4: *Shot size is a function of the distance between the camera and actors,as well as the camera focal length,as seen in (a). (b) shows the horizontal placement of actors in a composition is expressed in screen coordinates.*

Default *Screen* values are used to describe symetric compositions where subjects are evenly distributed from left to right. Non-default *Screen* values are used to describe asymetric compositions, e.g. taking into account head room and look room or the rules of thirds. We can also describe unconventional framing to create unbalanced artistic composition or to show other visual elements in the scene.

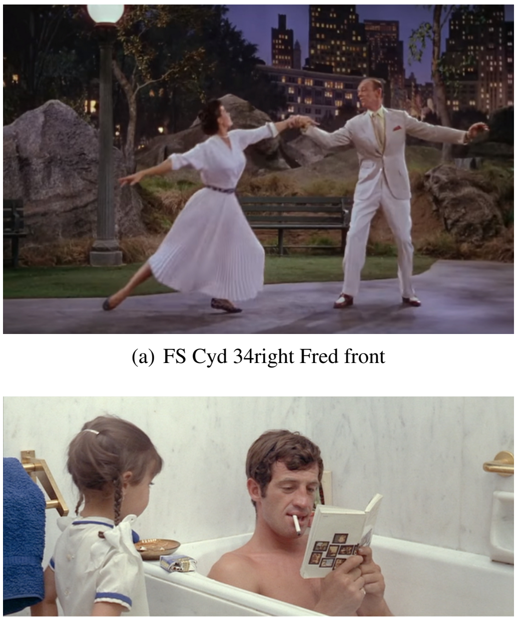
6. Shot Descriptions

A *shot* is a sequence of frames over a continuous time period. For a cohesive and coherent narration the individual shots have to be joined in a manner so as to allow the spectator to mentally recreate the story [Cut16]. Transition describes the progression of a shot to the subsequent one or the starting and ending of a shot. In our model we include three of the most widely used transition techniques: *cut*, *dissolve* and *fade*. We use the simplest form of cut transition in which the two shots are played one after another. We also use the same notation to describe other types of cuts, such as cutaways in which shot A is followed by a intermittent shot with a different composition and then returns back to shot A. Dissolves and fades are used to describe the entry or exit of a shot in which the composition either slowly appears or disappears respectively.



(a) low angle ECU Girl 34left

Figure 5: *Single actor composition in* Prose Storyboard Language. *This frame from Brian De Palma's* Dressed To Kill*(1980) shows an extreme close up shot from a low angle.*



(a) FS Cyd 34right Fred front

(b) MS Girl 34backright Ferdinand front

Figure 6: *Two actor composition in* Prose Storyboard Language. *Compositions from Vincente Minnelli's 1953 musical,* The Band Wagon *(a) and Jean-Luc Godard's 1965 French New Wave film,* Pierrot Le Fou *(b) feature two actors in a frame at different sizes.*

As we enter the shot the initial composition can be of two types. It can either be a 'static' composition in which the actors are not performing any action or we can transition into action. In the former case, we describe the composition in the first frame of the shot. In the latter, we describe the composition in relation to the action that is being performed (figure 1). This is described under the non-terminal *whileEvent* which describes the first composition of the shot *while* the actors are performing an *Action*. As we transition into action, the composition can also be accompanied by a camera movement. In Fig. 11 we transition directly into an actor action with camera movement. Here the camera tracks with the actor thereby resulting in maintaining a single composition in the entire action shot with no developments.

The key feature of the language is that all shots are self contained entities. Starting from a transition to the final composition, including camera and script actions, each prose storyboard sentence is independent of the previous or the next shot. A shot description can always be written and read without requiring knowledge from the previous or next shot in a movie.

Based on the taxonomy of shots proposed by Thomson and Bowen [TB09b], the prose storyboard language allows to describe precisely three main categories of shots :

- A simple shot is taken with a camera that does not move or turn. Any change in the composition is from the movements of the actors in relation to the camera.
- A complex shot is taken with a camera that has movements around a fixed point such as pan, tilt and zoom. We introduce camera actions pan and zoom to describe such movements. Thus the camera can pan left and right, up and down (as in a tilt) and zoom in and out.
- A developing shot is taken with a moving camera. We introduce two camera actions (dolly and crane) to describe these shots. Pan and zoom are allowed during dolly and crane movements thereby creating interesting visual effects.

Some shots consist of a single composition from beginning to end. In many cases, however, the initial composition is followed by a number of *developments*. They are of two types, *continuation* or *recomposition*. A *continuation* is the PSL description of actions, either actor or camera, that do not lead to a change in composition from the previous one. It can be from an *action* such as speaking or looking that is important to the screenplay to mention but that does not cause a change in composition. These are described under the *event* non-terminal. Or the *continuation* can be due to a *follow event* in which the subjects start to perform an action in the *cue* and the camera moves with the subjects and tracks their movements so as to maintain the composition. This camera action is categorized under the *camera with* non-terminal.

The second type of *development* is the *recomposition* where there is, as the name suggests, a change in composition. This change can either be due to an action performed by the *subject* or by the camera or both but these two movements are not synchronized as in *camera with*. These camera actions are categorized under the *camera to* non-terminal. In the case of a change in composition only because of actor movements such as in simple shots (figure 9), the camera "holds" to the next composition.

After the initial composition, there can be any number of developments of either type. To the best of our knowledge, the prose storyboard language is the first description framework that correctly describes developing shots of arbitrary length and complexity.

For each of the three cases we propose a simplified model of a shot which consists of a sequence of compositions, cues and screen events. Cues are actor movement which trigger camera movements. They are an important construct in classical movie-making where camera movements are frequently motivated by the story. Cues are optional, which makes it possible to also describe unmotivated camera movements preceding the action, or descriptive camera movements not related to actor movements. Screen events can be actions of the camera relative to the actors, or actions of the actors relative to the camera. Screen events come in two main categories - those which change the composition (*events-to-composition*) and those which maintain the composition (*events-with-composition*).

In our model, we can be in only one of four different states:

1. Camera does not move and composition does not change.
2. Camera does not move and composition changes due to actor movements.
3. Camera moves and composition does not change
4. Camera moves and composition changes.

In case (1), the shot can be described with a single composition. In case (2), we introduce the special verb *hold* to indicate that the camera remains static while the actors move, leading to a new composition. In case (3), we use the generic construction *CameraWith* (pan with, dolly with, crane with) to indicate how the camera moves to maintain the composition. In case (4), we use the generic construction *CameraTo* to indicate how the cameras moves to change the composition.

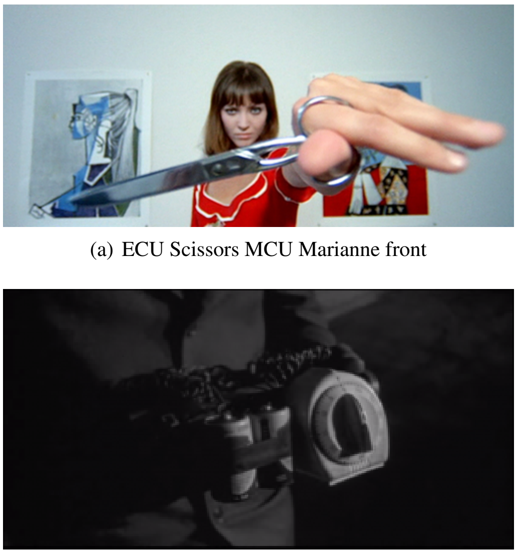
7. Annotation results

As a validation of the proposed language, we have manually annotated extended scenes from existing movies covering many different styles and periods.

7.1. Process of annotation

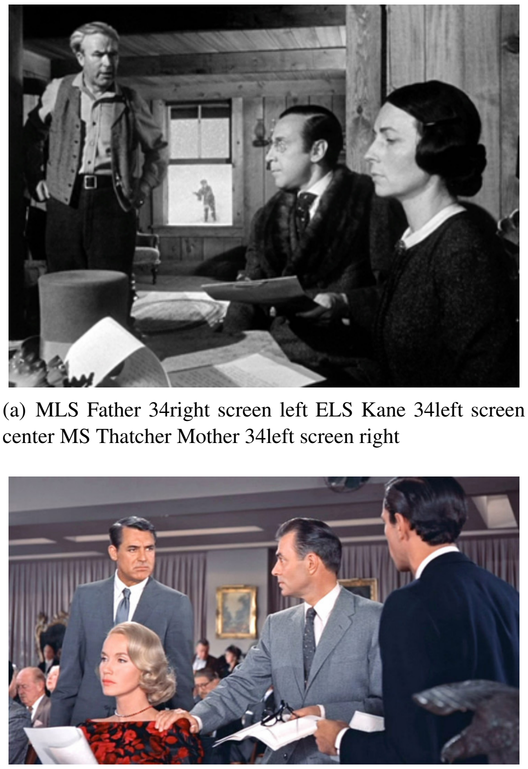
We start the annotation process by viewing the scenes multiple times. The scene is then divided into its consisting shots which we describe in Prose Storyboard Language. Each of these sentences is matched to their corresponding keyframe in the shot via timecodes. For example, the first frame of the shot matches the initial composition in PSL. The time code for this keyframe is noted. As the shot progresses, we make a note of the subsequent compositions and their time codes. This list of PSL sentences with time codes for keyframes is written as subtitles in a word processor and saved as SubRip(.srt) files that can be played with the annotated scene. To make the PSL sentences easier to read they are generally broken down into fragments that start with a 'from' composition and the time codes for the duration this composition lasts. Then the next PSL fragment contains the action that either the camera or actor or both perform that changes this initial composition. After this, we describe the 'to' composition that the previously mentioned action leads to. All these fragments are accompanied by their time codes. The output of this process of annotation is a time-coded PSL description of the scene in a .srt file.

These PSL sentences are then parsed using the PSL grammar described in the AND/OR tree in Fig. 14. This parsing is done in Python using the Parsimonious toolkit. The parser is based on parsing expression grammars (PEGs) in which lexing and parsing are done at the same time. The key feature of PEGs is that it uses a prioritized choice operator "/" rather than an unordered operator "|". This means the order in which the choices are written is important (For example, in a rule 'A = a / b / c ', the parser first checks if the input matches 'a'. It only moves to the next choice if this fails). This prioritization removes ambiguity and ensures there is only one output parse tree for a given input. A reference


(a) ECU Scissors MCU Marianne front

(b) CU Sanchez hands front as hands hold Bomb

Figure 7: *Compositions in (a) and (b) include inanimate objects (scissors in Jean-Luc Godard's* Pierrot Le Fou *and a bomb in Orson Welles's* Touch Of Evil.


(a) MLS Father 34right screen left ELS Kane 34left screen center MS Thatcher Mother 34left screen right

(b) MCU Eve 34left MS Thornhill 34right Vandam 34left Leonard 34backleft

Figure 8: *Complex composition in (a) Orson Welles's 1941* Citizen Kane *and (b) Alfred Hitchcock's 1959* North By Northwest *show multiple actors at different distances from the camera. They are described from left to right with their sizes indicating their depth in the composition.*

Table 1: *Annotation results: For each movie, we give the total number of annotated shots, compositions and developments, together with a count of the main categories of camera movement.*

Movie	**Shot**	**Composition**	**Development**							
			Continuation			Recomposition				
			Pan	Dolly	Crane	Hold	Zoom	Pan	Dolly	Crane
Back To The Future	41	69	-	-	-	12	1	6	10	-
North By Northwest	133	209	2	7	-	49	1	8	16	-
Touch Of Evil	1	40	-	2	1	6	-	1	18	13
Rope	2	12	-	1	-	-	-	3	7	-
Total	177	330	2	10	1	67	2	18	51	13

implementation of the PSL parser is freely available at https://gitlab.inria.fr/vmurukut/psl together with all PSL sentences mentioned in the paper.

7.2. Annotation results

We annotated scenes extracted from four movies : Back To The Future by Robert Zemekis, Rope and North By Northwest by Alfred Hitchcock, and Touch Of Evil by Orson Welles. In each case, we give the original screenplay, the movie subtitled with a complete PSL description of all compositions and developments, and a storyboard with one keyframe per composition or development. Annotation results are summarized in Table 1 and can be found in the accompanying material available online at https://team.inria.fr/anima/prose-storyboard-language/. With 177 shots and 330 compositions, they constitute an informal validation of the expressivity and generality of the language, as well as an illustration of good practices for precisely annotating movie shots using the language.

The cafe scene in Back To The future (1985) is an interior scene with a combination of action and dialogues involving 8 main characters. Scene elements for this movie are enumerated in Fig. 16 and appended to the PSL grammar for annotation. We are making the prose storyboard for all 41 shots in the scene available for future reference. Rope (1948), a single shot movie by Alfred Hitchcock, also shows action and dialogue between 8 characters, this time using elaborate blocking and camera movements rather than employing cuts. Scene elements for this movie are enumerated in Fig. 17. This is challenging example for annotation, and we show examples from two extended sequences fully annotated with PSL. Results are shown in Fig. 20 and 21.

We also annotated the crop duster scene from North By Northwest (1959) to highlight the versatility of the language in describing a complex scene in an outdoor environment involving many non human elements. In that scene the intent of the pilot is personified in the movements of the plane. We annotated all 133 shots in this virtuoso scene with their prose storyboards, to illustrate the variety of shots used in this mostly silent scene. Scene elements for this movie are enumerated in Fig. 18. Finally we annotated the long opening shot from Orson Welles's Touch Of Evil (1958) which shows a wide variety of camera movements interlaced with meticulously planned choreography for the characters resulting in a rich and dynamic visual composition. Scene elements for this movie are enumerated in Fig. 19. Despite the complexity of these scenes, they show that the prose storyboard is fairly simple to read and easy to generate.

8. Conclusion

We have presented a language for describing the spatial and temporal structure of movies with arbitrarily complex shots. The language can be extended in many ways, e.g. by taking into account lens choices, depth-of-field and lighting, and diegetic sound including speech. Future work will be devoted to the dual problems of automatically generating movies from prose storyboards in machinima environments, and automatically describing shots in existing movies. We are also planning to extend our framework for the case of stereoscopic movies, where image composition needs to be extended to include the depth and disparity of subjects in the composition. It would also be interesting to extend the language even further for the case of panoramic video and immersive virtual reality movies. At this stage, we believe that the proposed language can be useful to extend existing approaches in intelligent cinematography and editing towards more expressive strategies and idioms, and to bridge the gap between real and virtual movie-making.

References

[ASSN03] Akerberg O., Svensson H., Schulz B., Nugues P.: Carsim: An automatic 3d text-to-scene conversion system applied to road accident reports. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 2* (2003), EACL '03, pp. 191–194. 2

[BBAG08] Balas D., Brom C., Abonyi A., Gemrot J.: Hierarchical petri nets for story plots featuring virtual humans. In *AIIDE* (2008). 3

[Bor98] Bordwell D.: *On the History of Film Style*. Harvard University Press, 1998. 2

[Bra97] Brand M.: The "inverse hollywood problem": From video to scripts and storyboards via causal analysis. In *AAAI/IAAI* (1997), pp. 132–137. 1

[BS16] Braga P. H. C., Silveira I. F.: Slap: Storyboard language for animation programming. *IEEE Latin America Transactions 14*, 12 (Dec 2016), 4821–4826. 3

[BvKR01] Blackwell L., von Konsky B., Robey M.: Petri net script: a visual language for describing action, behaviour and plot. In *Australasian conference on Computer science* (2001), ACSC '01. 3

[CAwH*96] CHRISTIANSON D. B., ANDERSON S. E., WEI HE L., SALESIN D. H., WELD D. S., COHEN M. F.: Declarative camera control for automatic cinematography. In *AAAI* (1996). 1, 2

[Cha12] CHAO C.: Timing multimodal turn-taking for human-robot cooperation. In *Proceedings of the 14th ACM international conference on Multimodal interaction* (New York, NY, USA, 2012), ICMI '12, ACM, pp. 309–312. 3

[CO06] CHRISTIE M., OLIVIER P.: Camera control for computer graphics. In *Eurographics State of the Art Reports* (2006), Eurographics 2006, Blackwell. 1

[Cut16] CUTTING J. E.: Narrative theory and the dynamics of popular movies. In *Psychonomic Bulletin & Review* (2016). 4

[DMR05] DONY R., MATEER J., ROBINSON J.: Techniques for automated reverse storyboarding. *IEE Journal of Vision, Image and Signal Processing 152*, 4 (2005), 425–436. 1

[FF06] FRIEDMAN D., FELDMAN Y. A.: Automated cinematic reasoning about camera behavior. *Expert Syst. Appl. 30*, 4 (May 2006), 694–704. 2

[GCLR15] GALVANE Q., CHRISTIE M., LINO C., RONFARD R.: Camera-on-rails: Automated Computation of Constrained Camera Paths. In *ACM SIGGRAPH Conference on Motion in Games* (Paris, France, Nov. 2015), ACM, pp. 151–157. 1, 2

[GCR*13] GALVANE Q., CHRISTIE M., RONFARD R., LIM C.-K., CANI M.-P.: Steering Behaviors for Autonomous Cameras. In *MIG 2013 - ACM SIGGRAPH conference on Motion in Games* (Dublin, Ireland, Nov. 2013), MIG '13 Proceedings of Motion on Games, ACM, pp. 93–102. 1

[GCSS06] GOLDMAN D. B., CURLESS B., SALESIN D., SEITZ S. M.: Schematic storyboarding for video visualization and editing. *ACM Trans. Graph. 25*, 3 (2006), 862–871. 1

[GLC*18] GALVANE Q., LINO C., CHRISTIE M., FLEUREAU J., SERVANT F., TARIOLLE F.-L., GUILLOTEL P.: Directing cinematographic drones. *ACM Trans. Graph. 37*, 3 (July 2018), 34:1–34:18. 3

[GR13] GANDHI V., RONFARD R.: Detecting and naming actors in movies using generative appearance models. In *CVPR* (2013). 1

[GR15] GANDHI V., RONFARD R.: A Computational Framework for Vertical Video Editing. In *4th Workshop on Intelligent Camera Control, Cinematography and Editing* (Zurich, Switzerland, May 2015), Eurographics, Eurographics Association, pp. 31–37. 3

[GRCS14] GALVANE Q., RONFARD R., CHRISTIE M., SZILAS N.: Narrative-Driven Camera Control for Cinematic Replay of Computer Games. In *MIG'14 - 7th International Conference on Motion in Games* (Los Angeles, United States, Nov. 2014), ACM, pp. 109–117. 1, 2, 3

[GRG14] GANDHI V., RONFARD R., GLEICHER M.: Multi-Clip Video Editing from a Single Viewpoint. In *CVMP 2014 - European Conference on Visual Media Production* (London, United Kingdom, Nov. 2014), ACM, p. Article No. 9. 3

[HCS96] HE L.-W., COHEN M. F., SALESIN D. H.: The virtual cinematographer: a paradigm for automatic real-time camera control and directing. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques* (New York, NY, USA, 1996), SIGGRAPH '96, ACM, pp. 217–224. 1

[JY05] JHALA A., YOUNG R. M.: A discourse planning approach to cinematic camera control for narratives in virtual environments. In *AAAI* (2005). 2

[JY06] JHALA A., YOUNG R. M.: Representational requirements for a plan based approach to automated camera control. In *AIIDE'06* (2006), pp. 36–41. 1, 2

[LCL18] LOUARN A., CHRISTIE M., LAMARCHE F.: Automated staging for virtual cinematography. In *Proceedings of the 11th Annual International Conference on Motion, Interaction, and Games, MIG 2018, Limassol, Cyprus, November 08-10, 2018* (2018), pp. 4:1–4:10. 3

[LDTA17] LEAKE M., DAVIS A., TRUONG A., AGRAWALA M.: Computational video editing for dialogue-driven scenes. *ACM Trans. Graph. 36*, 4 (July 2017), 130:1–130:14. 2

[LG90] LITTLE T. D. C., GHAFOOR A.: Synchronization and storage models for multimedia objects. *IEEE Journal on Selected Areas in Communications 8* (1990), 413–427. 3

[MJSB11] MARKOWITZ D., JR. J. T. K., SHOULSON A., BADLER N. I.: Intelligent camera control using behavior trees. In *MIG* (2011), pp. 156–167. 2

[MRR98] MAGALHAES L. P., RAPOSO A. B., RICARTE I. L.: Animation modeling with petri nets. *Computers and Graphics 22*, 6 (1998), 735 – 743. 3

[Mur86] MURCH W.: *In the blink of an eye*. Silman-James Press, 1986. 3

[MVW*18] MARTI M., VIELI J., WITOŃ W., SANGHRAJKA R., INVERSINI D., WOTRUBA D., SIMO I., SCHRIBER S., KAPADIA M., GROSS M.: Cardinal: Computer assisted authoring of movie scripts. In *23rd International Conference on Intelligent User Interfaces* (2018), IUI '18, pp. 509–519. 2

[Ond04] ONDAATJE M.: *The Conversations: Walter Murch and the Art of Film Editing*. Random House, 2004. 3

[ORN09] O'NEILL B., RIEDL M. O., NITSCHE M.: Towards intelligent authoring tools for machinima creation. In *CHI Extended Abstracts* (2009), pp. 4639–4644. 2

[Pro08] PROFERES N.: *Film Directing Fundamentals - See your film before shooting it*. Focal Press, 2008. 2

[RKV*09] RIJSSELBERGEN D. V., KEER B. V. D., VERWAEST M., MANNENS E., DE WALLE R. V.: Movie script markup language. In *ACM Symposium on Document Engineering* (2009), pp. 161–170. 2, 3

[Sal06] SALT B.: *Moving Into Pictures*. Starword, 2006. 4

[Sal09] SALT B.: *Film Style and Technology: History and Analysis (3 ed.)*. Starword, 2009. 2

[SMAY03] SHEN J., MIYAZAKI S., AOKI T., YASUDA H.: Intelligent digital filmmaker dmp. In *ICCIMA* (2003). 2

[SY06] SEVERSKY L. M., YIN L.: Real-time automatic 3d scene generation from natural language voice and text descriptions. In *Proceedings of the 14th ACM International Conference on Multimedia* (2006), MM '06, pp. 61–64. 2

[TB09a] THOMPSON R., BOWEN C.: *Grammar of the Edit*. Focal Press, 2009. 1

[TB09b] THOMPSON R., BOWEN C.: *Grammar of the Shot*. Focal Press, 2009. 1, 4, 5

[VHL*13] VOGT J., HAESEN M., LUYTEN K., CONINX K., MEIER A.: Timisto: A technique to extract usage sequences from storyboards. In *Proceedings of the 5th ACM SIGCHI Symposium on Engineering Interactive Computing Systems* (2013), EICS '13, pp. 113–118. 3

[WPRC18] WU H.-Y., PALÙ F., RANON R., CHRISTIE M.: Thinking like a director: Film editing patterns for virtual cinematographic storytelling. *ACM Trans. Multimedia Comput. Commun. Appl. 14*, 4 (Oct. 2018), 81:1–81:22. 3

[Yan13] YANNOPOULOS A.: DirectorNotation: Artistic and technological system for professional film directing. *J. Comput. Cult. Herit. 6*, 1 (Apr. 2013), 2:1–2:34. 3

[YB08] YE P., BALDWIN T.: Towards automatic animated storyboarding. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 1* (2008), AAAI'08, pp. 578–583. 2

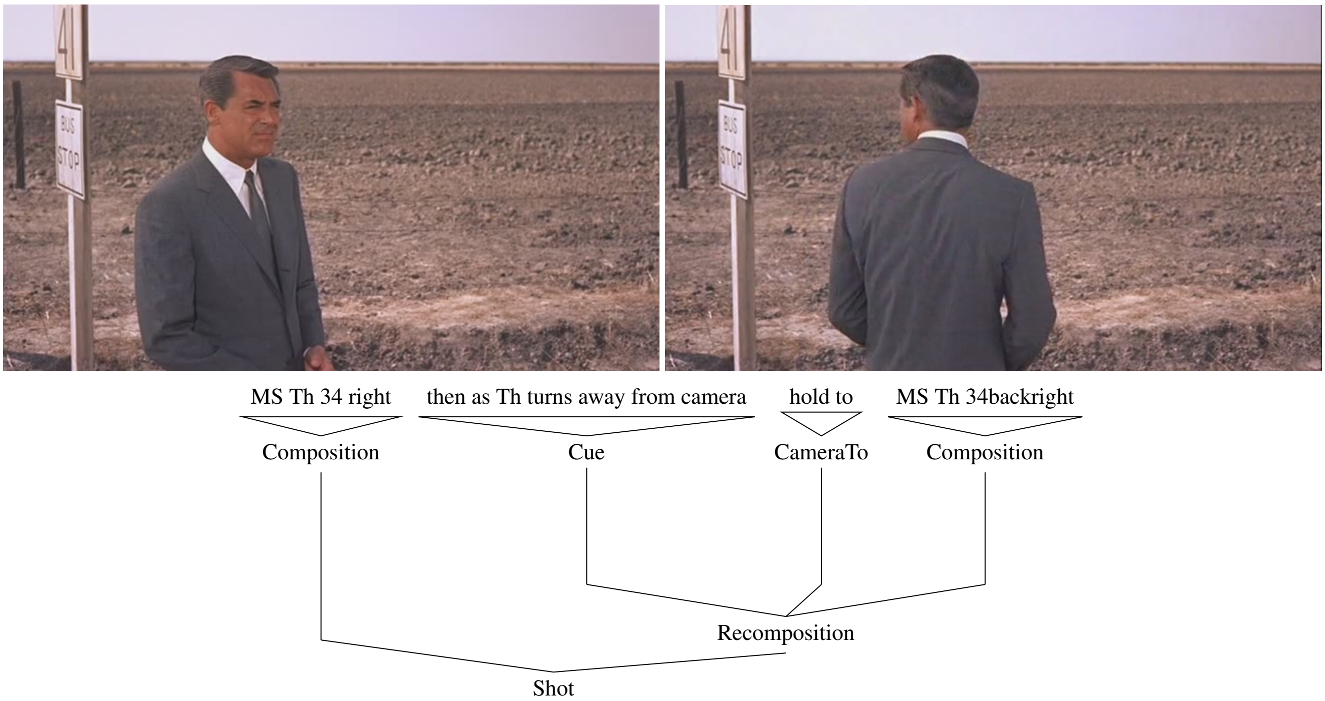


Figure 9: *Simple shot: with actor movement in* North By Northwest

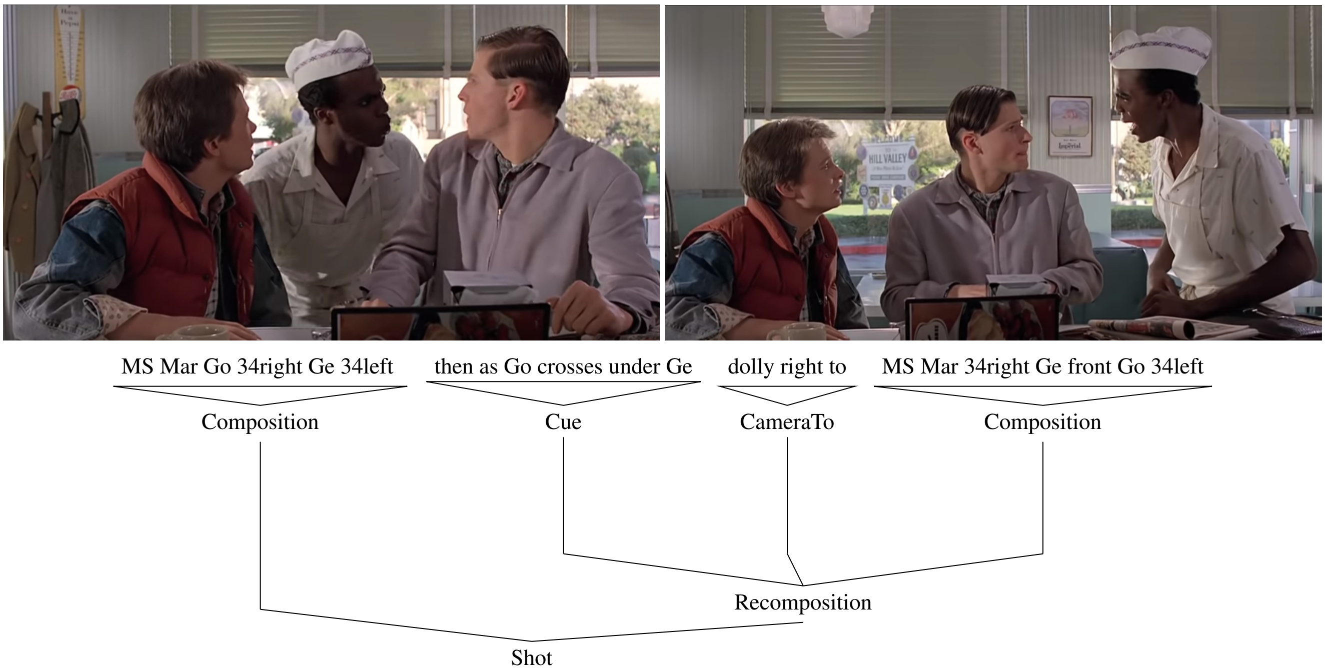

MS Mar Go 34right Ge 34left
then as Go crosses under Ge
dolly right to
MS Mar 34right Ge front Go 34left
Composition
Cue
CameraTo
Composition
Recomposition
Shot

Figure 10: *Developing shot in* Back To The Future

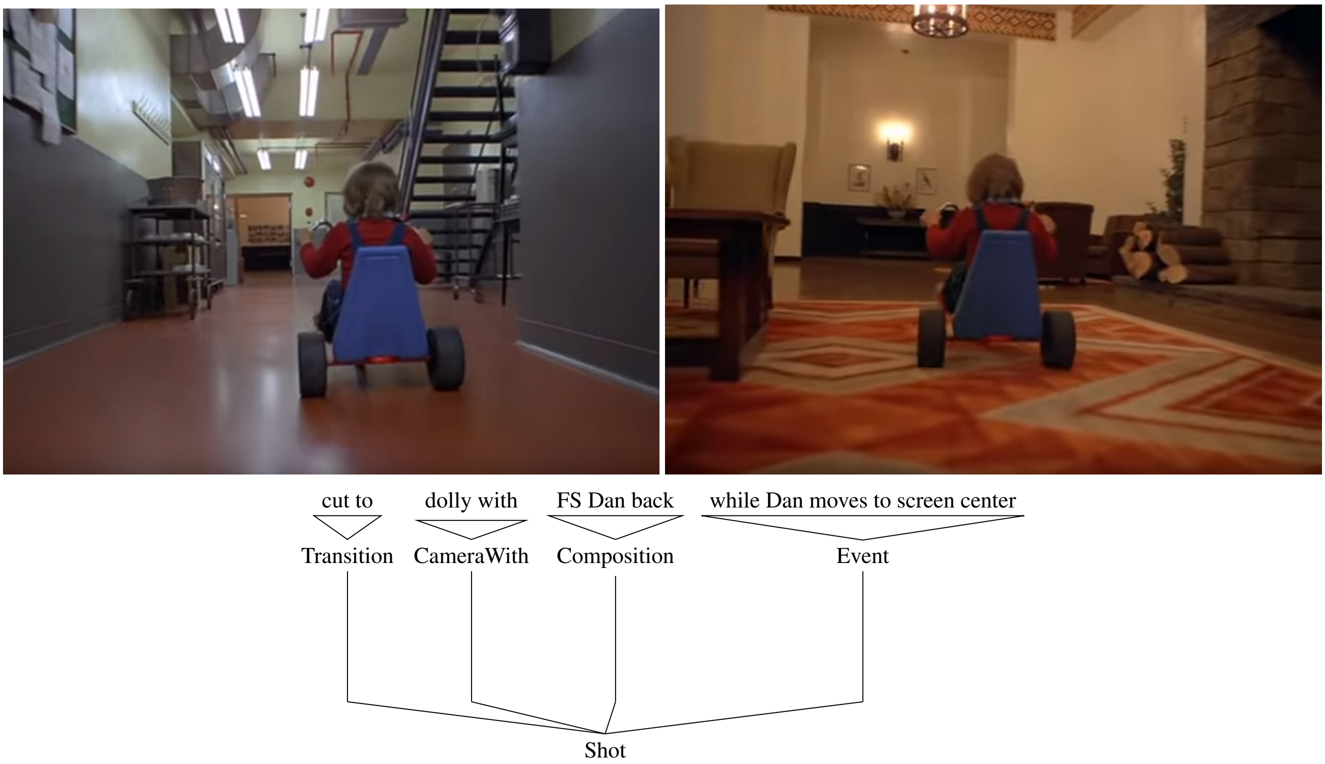


Figure 11: *Developing shot: Dolly with actor in* The Shining

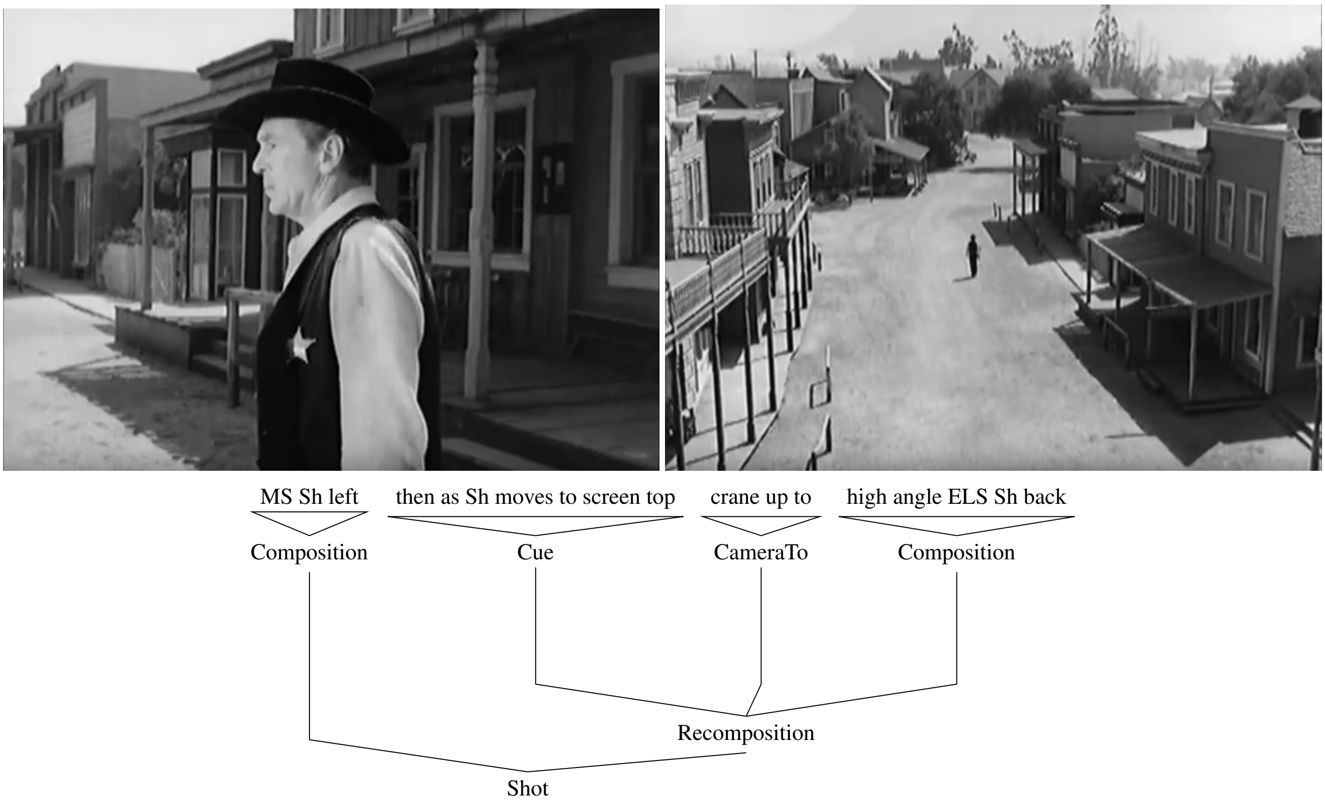


Figure 12: *developing shot: Crane up in* High Noon

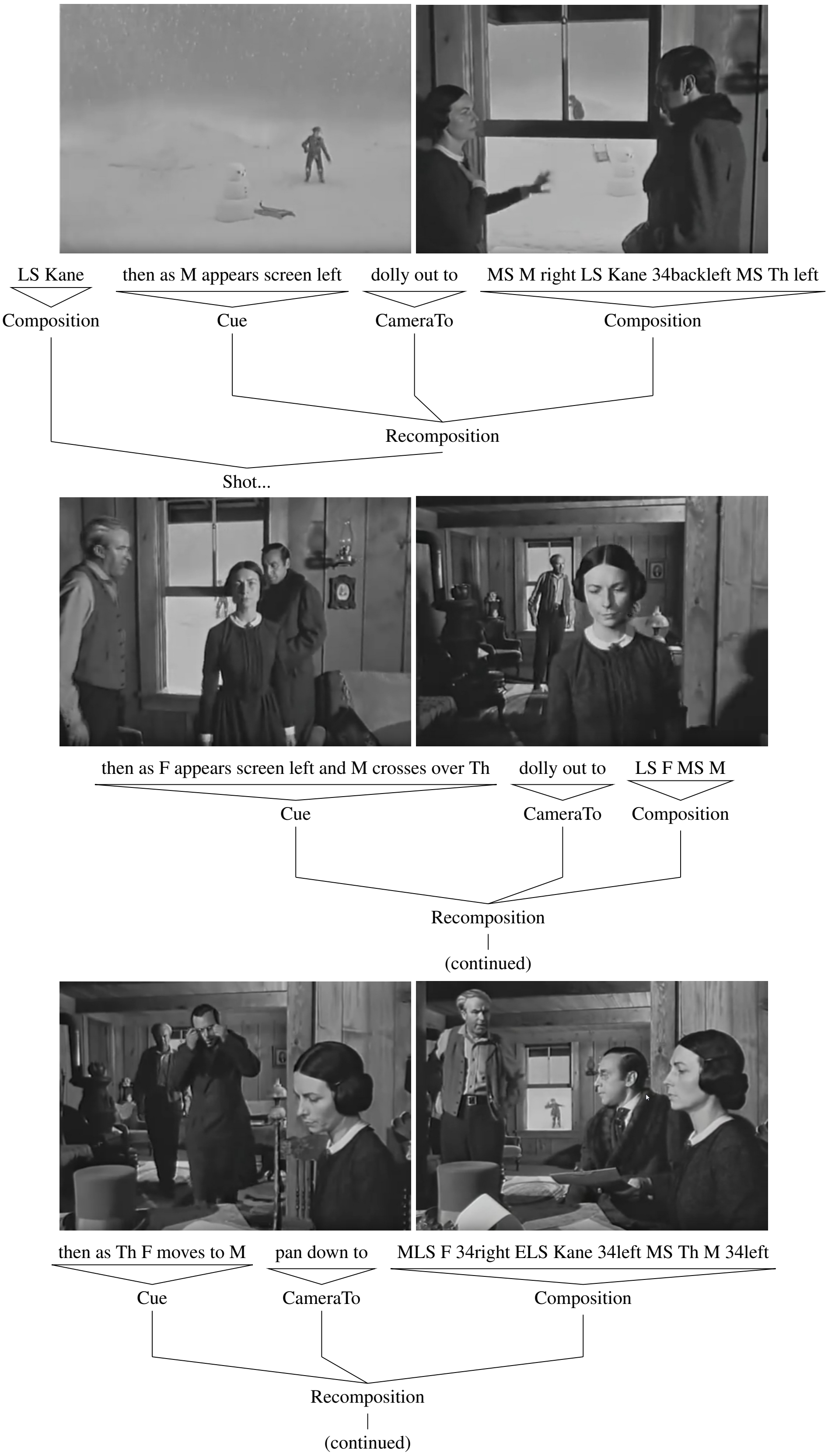


Figure 13: *Developing shot with multiple actors in* Citizen Kane

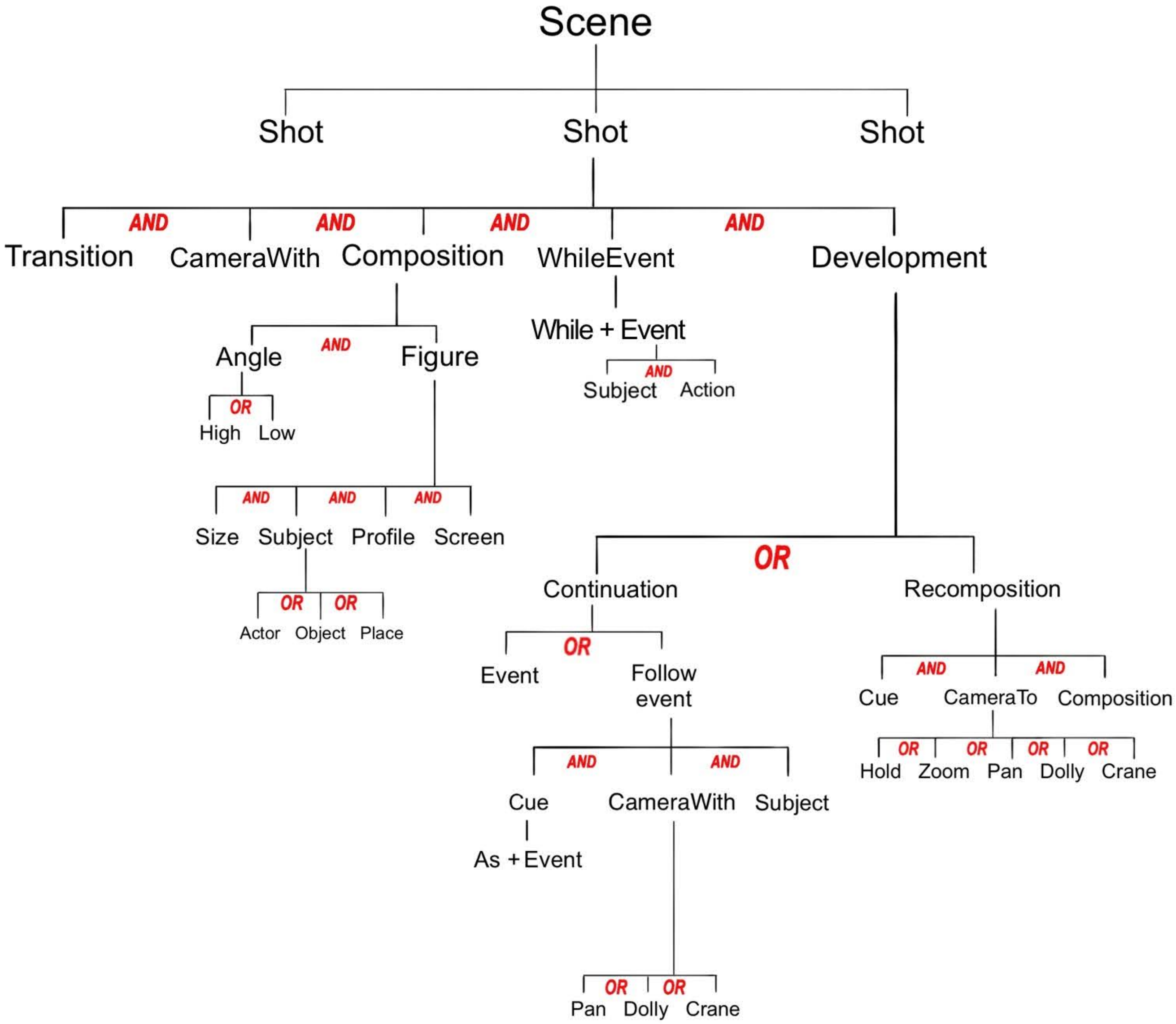


Figure 14: *AND-OR tree representation of the Prose Storyboard Language grammar.*

```
Scene          = Shot*
Shot           = Transition? _ CameraWith? _ Composition? _ WhileEvent?
                  _ (Development?)*
Transition     = ("cut to" / "dissolve to" / "fade in to")
Development    = "then"? _ (Recomposition / Continuation)
Continuation   = Event / FollowEvent
Recomposition  = Cue? _ CameraTo _ Composition
WhileEvent     = "while"_ Event
FollowEvent    = Cue? _ CameraWith _ Agent _ ("and"? _ Agent)*
Cue            = "as" _ Event _ ("and"? _ Event)*
Composition    = (Angle? _ Figure)*
Figure         = Size? _ Subject _ Profile? _ Screen?
Agent          = (Actor / Object) _ (Actor / Object)*
Subject        = Actor / Object / Place
Angle          = "low angle" / "high angle"
Size           = "ECU" / "CU" / "MCU" / "MS" / "MLS" / "FS" / "LS" / "ELS"
Profile        = "left" / "right" / "front" / "back" / "34left" / "34right"
                  / "34backleft" / "34backright"
Screen         = "screen" _ ("top" / "bottom")? _ ("left" / "center"/ "right")?
CameraWith     = Speed? _ (Pan / Dolly / Crane)  _ "with"
CameraTo       = (Hold /(Speed? _ (Pan / Dolly / Crane / Zoom))) _ "to"
Hold           = "hold"
Pan            = "pan"   _ ("left" / "right" / "up" / "down")?
Dolly          = "dolly" _ ("left" / "right" / "in" / "out")?
Crane          = "crane"  _ ("up" / "down") _ ("left"/"right")?
Zoom           = "zoom" _ ("in" / "out")
Speed          = "slow" / "quick"
Enter          = "enters"  _ Screen? _ Place?
Exit           = "exits" _ Screen? _ Place?
Look           = "looks" _ "at"? _ (Subject / Screen)
Move           = "moves" _ "to" _ (Subject / Screen )
Speak          = ("speaks" / ("says" _ String)) _ ("to" _ Subject)?
Use            = "uses" _ Object
Cross          = "crosses" _ ("over" / "under") _ Subject
Touch          = "touches" _ Subject
React          = "reacts to" _ Subject
Turn           = "turns" _ ("left" / "right"
                  / "towards camera" / "away from camera")
Stop           = "stops" _ ("at" / "near") _ Subject
Appear         = "appears" _ (Screen / ("from behind" _ Subject))
Disappear      = "disappears" _ (Screen/ ("behind" _ Subject))
Action         = Enter / Exit / Look / Move / Speak / Use / Cross / Touch
                  / React / Turn / Stop / Appear / Disappear
Event          = Agent _ Action
```

Figure 15: *Grammar of the prose storyboard language in the Parsing Expression Grammar (PEG) format.*

```
Actor          = "Marty" / "George" / "Biff" / "Lou" / "Goldie"
                  / "Match" / "Skinhead" / "hands" / "3D"
Object         = "Coffee"/ "Bar" / "Car"
Place          = "Cafe"
```

Figure 16: *Script elements for Back To The Future.*

```
Actor          = "Brandon" / "Philip" / "Atwater" / "Janet" /
                   "Kentley" / "Kenneth" / "Rupert"  / "Wilson"
Object         = "Glass"
Place          = "Salon" / "Dining room" / "Kitchen"
```

Figure 17: *Script elements for Rope.*

```
Actor        = "Thornhill" / "MBS" / "Plane" / "TD1" / "TD2" / "Farmer" /
               "BC Driver" / "BC Woman" / "BC Man"
Object       = "Bus" / "White car" / "Limo" / "Truck" / "Blue car" /
               "Green bus" /"Bluewhite car" / "Oil truck" / "Pickup"
                / "Brown car"
Place        = "Arid plot 1" / "Arid plot 2" / "Arid plot 3" /
               "Corn field" / "Highway" / "Dirt road"
```

Figure 18: *Script elements for North By Northwest.*

```
Actor        = "Mike" / "Susan" / "Linnekar" / "Blonde" / "Sanchez" /
                "Immigration official" / "Customs official"
Object       = "Car" / "Bomb" / "Building" /"Checkpost"
Place        =  "Border control" / "Main street" / "Parking lot"
                / "Left side street"
```

Figure 19: *Script elements for Touch Of Evil.*

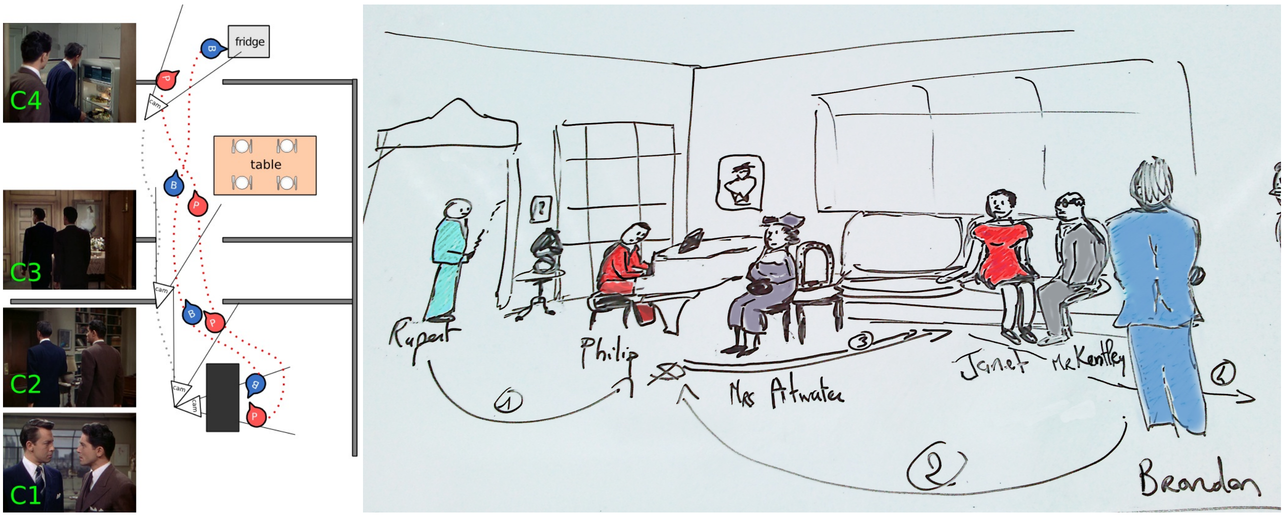


Figure 20: *Sketch storyboards for two sequences in Alfred Hitchcock's Rope, see Fig.21 below.*

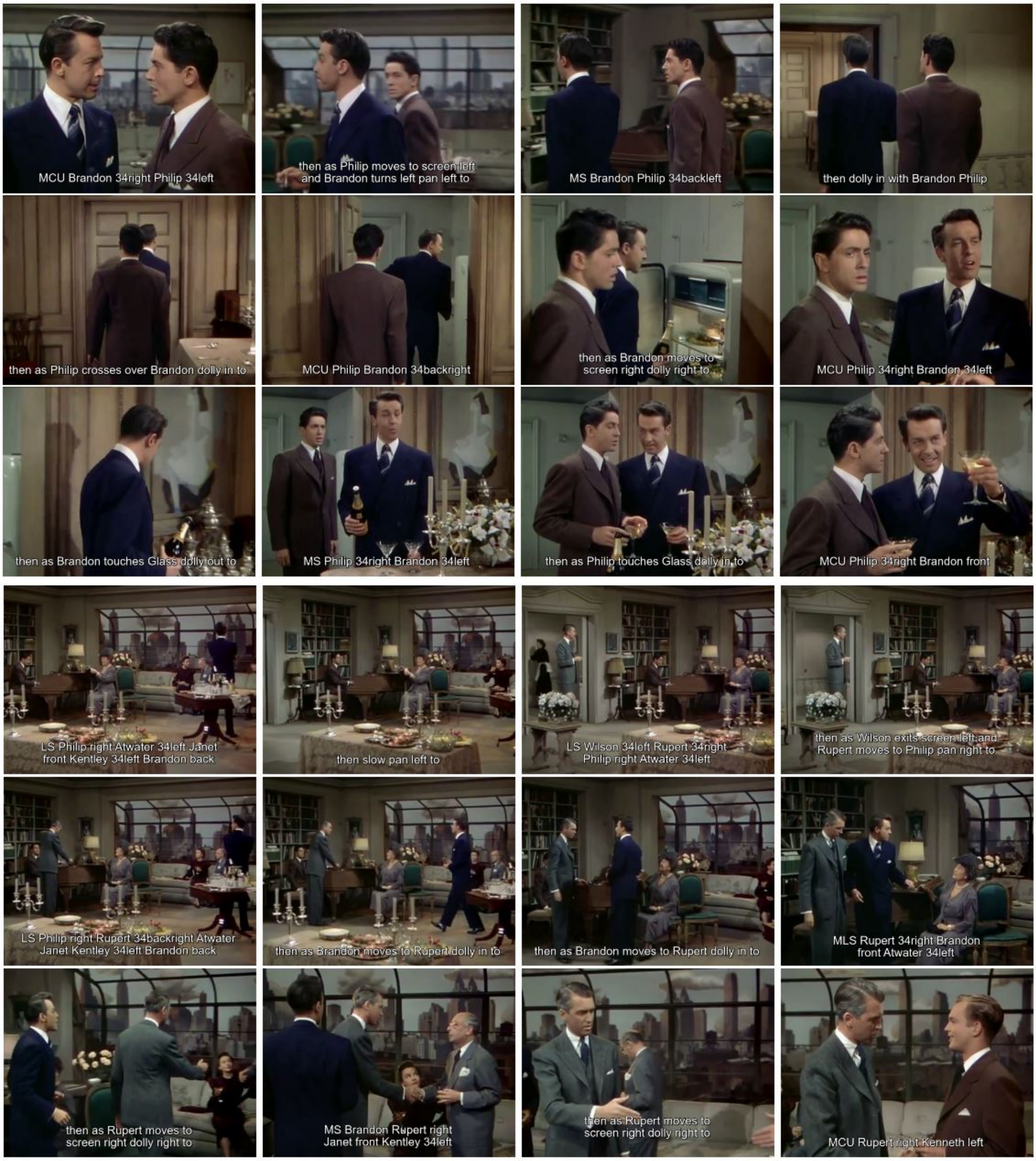

MCU Brandon 34right Philip 34left
then as Philip moves to screen left and Brandon turns left pan left to
MS Brandon Philip 34backleft
then dolly in with Brandon Philip
then as Philip crosses over Brandon dolly in to
MCU Philip Brandon 34backright
then as Brandon moves to screen right dolly right to
MCU Philip 34right Brandon 34left
then as Brandon touches Glass dolly out to
MS Philip 34right Brandon 34left
then as Philip touches Glass dolly in to
MCU Philip 34right Brandon front
LS Philip right Atwater 34left Janet front Kentley 34left Brandon back
then slow pan left to
LS Wilson 34left Rupert 34right Philip right Atwater 34left
then as Wilson exits screen left and Rupert moves to Philip pan right to
LS Philip right Rupert 34backright Atwater Janet Kentley 34left Brandon back
then as Brandon moves to Rupert dolly in to
then as Brandon moves to Rupert dolly in to
MLS Rupert 34right Brandon front Atwater 34left
then as Rupert moves to screen right dolly right to
MS Brandon Rupert right Janet front Kentley 34left
then as Rupert moves to screen right dolly right to
MCU Rupert right Kenneth left

Figure 21: *Prose storyboard language annotations of two extended sequences from the movie* Rope. *Top three rows: First sequence from 06:55 to 08:22. Bottom three rows: Second sequence from 10:00 to 12:00.*