Eurographics Workshop on Intelligent Cinematography and Editing (2017)
W. Bares, V. Gandhi, Q. Galvane, and R. Ronfard (Editors)

Five Challenges for Intelligent Cinematography and Editing

Remi Ronfard

Univ. Grenoble Alpes, Inria, Grenoble, France

Abstract

In this position paper, we propose five challenges for advancing the state of the art in intelligent cinematography and editing by taking advantage of the huge quantity of cinematographic data (movies) and metadata (movie scripts) available in digital formats. This suggests a data-driven approach to intelligent cinematography and editing, with at least five scientific bottlenecks that need to be carefully analyzed and resolved.we briefly describe them and suggest some possible avenues for future research in each of those new directions.

Categories and Subject Descriptors (according to ACM CCS): I.2.10 [Vision and Scene Understanding]: —I.3.3 [Computer Graphics]: —

1. Building a database of movie scenes

There have been several attempts in the past to build databases of movie scenes for the purpose of action recognition. One interesting line of research is to use movie scripts as a source of weak annotation, where action verbs in the script are used to identify actions in the movie. Cour et al. describe a method for aligning movies and scripts using telext subtitles and building a database of short movie clips described by action verbs [CT07, CJMT08]. Laptev et al. adopt a similar strategy for building the so-called "hollywood" dataset (HOHA). They automatically extract video segments that likely contain the actions mentioned in the aligned script [LMSR08]. Those segments are then verified manually. Gupta et al. use teletex transcriptions of sports broadcasts for building a database of sports actions [GSSD09]. Salway et al. use audio descriptions for creating a database of movies with a rich description of actions [AVA05, SLO07]. Rohrbach et al. align movies with audio descriptions to create a parallel corpus of over 68K sentences and video snippets from 94 HD movies. [RRTS15].

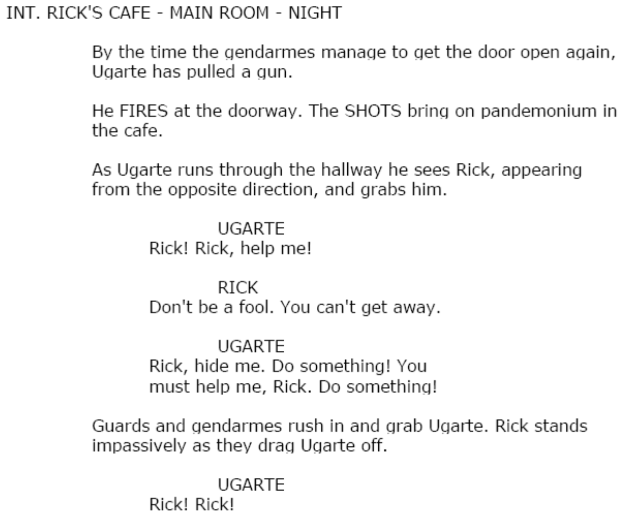
Such databases are useful for the purpose of action recognition, but are not sufficient for learning models of cinematography and film editing. For one thing, the number of action classes and examples per action class are usually small (the HOHA database contains 430 video segments labeled with 8 action classes). Furthermore, they do not preserve the structure of the movies into cinematographic scenes and shots. For the purpose of learning general models of cinematography and editing, a much larger number of movie scenes will be needed with a much more diverse set of actions and cinematographic styles. A movie generally contains in the order of a hundred scenes. Therefore, a complete alignment of one hundred movies with their scripts can be expected to yield a database of ten thousand movie scenes. This will require an intense effort from our community because the problem of detecting scene breaks in movies remains difficult in general.

A possible approach to this problem is to train scene break classifiers from labeled examples. Active learning methods should be used to refine classifiers using false negatives (scene breaks) and false positives (non scene breaks) collected by film experts. Another possible approach is to detect scene breaks as part of the script-to-movie alignment process. This will require a more general form of alignment where both the movie and the script are represented as tree structures (the movie contains scenes which contain shots which contain frames, the script contains scenes which contain actions and dialogues). The alignment should then be performed between trees, rather than sequences, and specific methods such as [HO82] should be used, rather than the commonly used digital time warping (DTW).

2. Breaking down scenes into shots

After collecting and annotating a large collection of movie scenes, we will have to confront two related methodological issues. The first issue is the size of the vocabulary of actions present in those scenes, which will likely be in be in the order of several thousand concepts. This makes the traditional approaches of learning action concepts one by one impractical and beyond the reach of the inteligent cinematography and editing community. The second issue is that the action labels present in the script are only a very rough and incomplete description of the action actually performed in screen. The art of mise-en-scene and acting consist primarily in translating the more abstract action concepts present in the script into the more concrete actions played to the camera. This is best illustrated by comparing the original screenplay for a short movie scene from the movie ’Casablanca" reproduced in Figure 1 with the actions per-

DOI: 10.2312/wiced.20171069



INT. RICK'S CAFE - MAIN ROOM - NIGHT

By the time the gendarmes manage to get the door open again, Ugarte has pulled a gun.

He FIRES at the doorway. The SHOTS bring on pandemonium in the cafe.

As Ugarte runs through the hallway he sees Rick, appearing from the opposite direction, and grabs him.

UGARTE
Rick! Rick, help me!

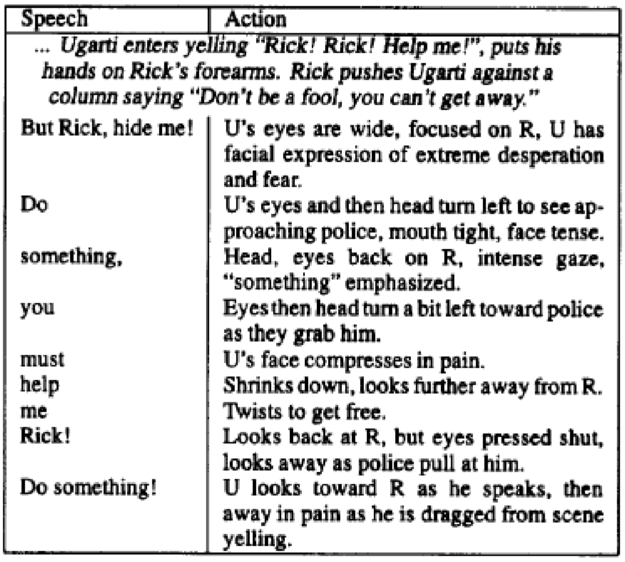
RICK
Don't be a fool. You can't get away.

UGARTE
Rick, hide me. Do something! You
must help me, Rick. Do something!

Guards and gendarmes rush in and grab Ugarte. Rick stands impassively as they drag Ugarte off.

UGARTE
Rick! Rick!

Figure 1: *A scene from the original screenplay of the movie Casablanca.*



Speech	Action
... Ugarti enters yelling "Rick! Rick! Help me!", puts his hands on Rick's forearms. Rick pushes Ugarti against a column saying "Don't be a fool, you can't get away."	
But Rick, hide me!	U's eyes are wide, focused on R, U has facial expression of extreme desperation and fear.
Do	U's eyes and then head turn left to see approaching police, mouth tight, face tense.
something,	Head, eyes back on R, intense gaze, "something" emphasized.
you	Eyes then head turn a bit left toward police as they grab him.
must	U's face compresses in pain.
help	Shrinks down, looks further away from R.
me	Twists to get free.
Rick!	Looks back at R, but eyes pressed shut, looks away as police pull at him.
Do something!	U looks toward R as he speaks, then away in pain as he is dragged from scene yelling.

Figure 2: *Shot-by-shot description (decoupage) of the same scene from the movie Casablanca as in Figure 1 shots 5 and 6 (reproduced from Loyall and Bates [LB97]). The screenplay contains only five actions: "Ugarte runs, sees Rick, grabs him, "guards rush in and grab Ugarte". The decoupage contains many more subtle interactions between dialogue, non verbal communication and physical actions. Can a statistical model of mise-en-scene be learned for translating the (hidden) screenplay actions into the (visible) movie actions ?*

formed by the actors in the movie , which are described in Figure 2 and Figure 3.

In order to overcome those difficulties, we believe it will be necessary to perform a shot-by shot annotation of each scene in the corpus. In some exceptional cases, such shot descriptions are available in the shape of a *decoupage* (continuity script) which can be automatically aligned to the movie scene [RTT03, Ron04]. In the more general case, the shot-by-shot annotation must be created from scratch, using controled vocabularies and formal description languages such as the *prose storyboard language* [RGB15], which has been shown to be expressive enough to describe movie shots and scenes with arbitrary complexity. Movie scenes contain on the order of twenty shots, which means that a collection of 10,000 scenes will comprise approximately 2 million shots. Clearly, this annotation cannot be obtained manually and future work is need to automate it at least partially.

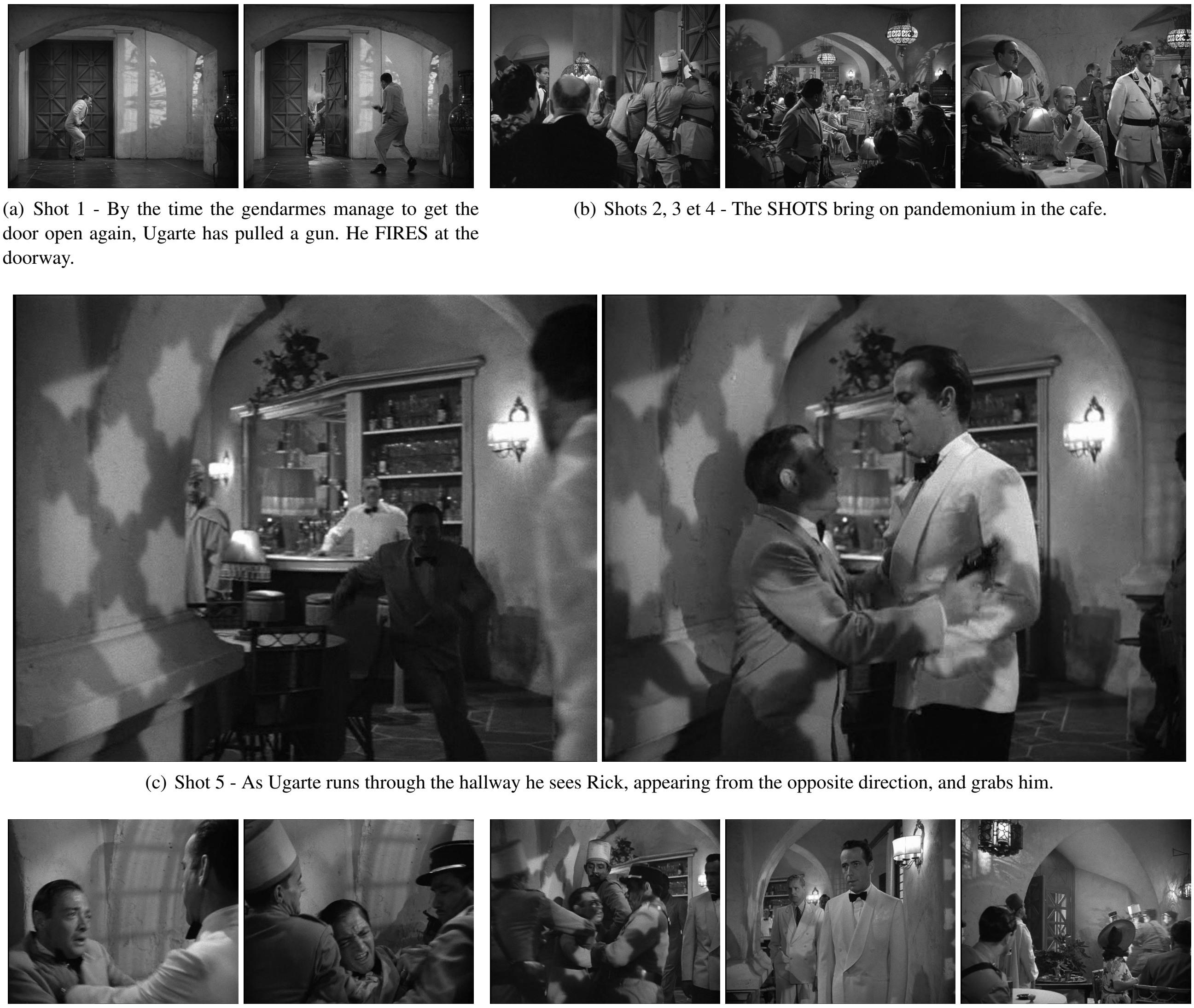
A promising approach will be to train conditional random fields (CRF) from examples of fully-described shots and to attempt to generalize to novel shots from the same movie or the same genre. Similar approaches have been proposed recently for describing still images using scene graphs [JKS*15] and we conjecture that they will generalize to the more difficult problem of describing movie shots using prose storyboards.

3. Recognizing actors and their actions

Using the temporal alignments between prose storyboards and movie shots will put us in a good position for learning to recognize movie actors and their actions, and to understand the different cinematographic and editing styles which are used to portray them. In previous work, we obtained good results in simultaneously learning models of actions and viewpoints using hidden Markov models [WBR07]. We therefore conjecture that similar approaches can be used for simultaneously learning models of movie actors and their actions (content) together with corresponding shot composition and editing (style). Marginalizing over style parameters, we can expect to obtain improved precision in the difficult task of human action recognition, which is notoriously hard in movies . Marginalizing over content parameters, we can expect to learn useful models of cinematography and editing styles, well beyond the current state of the art in the statistical analysis of film style [CDN10, CC15]. What makes this problem particularly challenging is the huge size of the vocabulary both in content and in style (to be compared with the 11 action categories and 8 view points learned by Weinland et al [WBR07]).

Luckily, recent advances in computer vision are making human body and face detection reliable enough that it becomes possible to reformulate the action recognition problem. Instead of asking the harder question - what is happening in this shot or scene ? we can now ask an easier question - what is this actor doing in this shot or scene ? Relying on actor body and face detection brings the additional advantage that we can describe the video in body coordinates, which are suitable representation for human actions and activities. In this context, a very promising approach for recognizing a large vocabulary of actions will be to learn semi-Markov conditional random field (SMCRF) models using variants of backpropagation [Col02, SWCS08].

Despite the spectacular recent progress in large-scale machine learning, we would like to argue that learning models of action and



(a) Shot 1 - By the time the gendarmes manage to get the door open again, Ugarte has pulled a gun. He FIRES at the doorway.

(b) Shots 2, 3 et 4 - The SHOTS bring on pandemonium in the cafe.

(c) Shot 5 - As Ugarte runs through the hallway he sees Rick, appearing from the opposite direction, and grabs him.

(d) Shot 6 - Quick dialogue between Ugarte and Rick. Guards and gendarmes rush in and grab Ugarte.

(e) Shots 7, 8 et 9 - Rick stands impassively as they drag Ugarte off.

Figure 3: *Keyframes from the movie 'Casablanca' corresponding to the scene scripted in in Figure 1. The scene was filmed and edited with nine different shots, elaborating on the much shorter action description present in the original screenplay. The translation from script to shot (decoupage) is a major component of film directing, involving actor direction as well cinematography and editing. Understanding the complexity of decoupage is a key challenge for intelligent cinematography and editing and requires a careful breakdown and analysis of classic scenes into shots.*

cinematography in a purely data-driven fashion, may not be sufficient. As a supplementary source of information, it would be useful to create synthetic examples, where the different parameters of cinematic styles, including blocking, lighting and camera framing, can be generated in a more systematic fashion.

This leads to the challenge of creating realistic simulations of movie scenes in 3-D animation. In previous work, we recreated one short scene from the movie 'back to the future' for the purpose of demonstrating the performance of our automatic film editing method [GRLC15] and comparing it to the actual editing in the movie [GRC15]. In related work, researchers have started to use game engines to reproduce movie scenes as part of the 'machinima' movement [KM05, Low05, Nit09]. Such techniques show great promise for generating variations in movie-making using virtual sets, actors and cameras, which leads us to our next challenge.

4. Reverse-engineering movie scenes

Starting from an example movie scene broken down into shots, together with a detailed screenplay describing the dramatic action and a prose storyboard describing the composition of each shot, we are in a good position for re-creating the scene in 3-D animation using the tools of machinima. Existing software tools such as Persona and Matinee in the Unreal Engine, facilitate the creation of such cinematic sequence using a combination of live interaction and scripted animation. The Source Recorder in Valve provides similar support. Open source game engines such as Panda3D or Blender can also be used tor replicate movie scenes using existing assets.

Machinima pioneer Michael Nitsche gives a very detailed account of the machinima reconstruction of the movie 'Casablanca' during two workshop at the University of Cambridge in 2002 et 2003 [NM04, Nit09]. Those reports clearly illustrate the promises and the limitations of existing machinima tools. Even today, the effort required to recreate a movie scene in machinima remains enormous, as we have experienced ourselves while re-creating the scene in 'Back to the future' in our lab. But the reward is also substantial, as that 3-D model make it possible to generate a large number of variations in style for the same content, and to uses them for learning invariant action recognition methods [dSGCP16].

Generating those 3-D scenes automatically is our next challenge. Currently, each step in the reconstruction requires laborious interaction - including set reconstruction, virtual actor modeling, retargeting of full body animation from motion capture databases, facial animation and lip-sync, synchronisation between actors, collision detection and physical simulation of the environment. In future work, it should be possible to use prose storyboards [RGB15] as a scripting language for automatically generating 3-D scenes in machinima.

We believe this is a realistic goal, much more so than the previous work of Loyall et Bates [LB93] or Ye and Baldwin [YB08] who attempted to create 3-D scenes directly from movie scripts, without the intermediate step of the prose storyboard. In this endeavour, Loyall et Bates proposed the HAP language [LB93], which uses the framework of behavior trees for scripting actions and reactions of virtual actors in response to their environment. Actor behaviors are computer programs with names (goals), parameters, sub-goals, pre-conditions and post-conditions. They can run sequentially or in parallel.

A promising direction for future research will be to build a probabilistic version of the HAP language, with probabilities computed from examples of real movie scenes. Such a language could be used to learn statistical models of actions and acting styles and to reuse them during machinima generation. This process of reverse-engineering movie scenes in terms of generative models would make it possible for our community to share large numbers of scenes with a variety of contents and styles, suitable for learning more realistic models of cinematography and editing, not limited to a single movie, director, era or genre.

5. Generating movie scenes

Given a large enough number of movie scenes and their reverse-engineered, 3-D animation versions, it becomes possible to reformulate the problem of cinematography and editing as a regression problem, in the way of recent attempts to translate video into text and vice versa using deep neural networks [SVL14, VRD*15].

We expect that such methods will eventually make it possible to generate movie scenes with the complexity of the 'Casablanca' example in Figure 3 on a much larger scale than is currently possible. This short movie scene is much more complex and compelling than all the previous work in intelligent cinematography and editing, which uses relatively simple, sometimes caricatural 3-D graphics and animation.

The promises and the challenges of the data-driven approach that we advocate are equally great. Each of the challenges will require a much needed collaboration between researchers in computer vision and computer graphics, knowledge engineers, film scholars and machine learning specialists. That may be the ultimate challenge for our community.

References

[AVA05] ANDREW A. S., VASSILIOU A., AHMAD K.: What happens in films? In *IEEE International Conference on Multimedia and Expo* (2005). 1

[CC15] CUTTING J., CANDAN A.: Shot durations, shot classes, and the increased pace of popular movies. *Projections: The journal for movies and mind 9*, 2 (2015), 40–62. 2

[CDN10] CUTTING, DELONG, NOTHELFER: Attention and the evolution of hollywood film. *Psychological Science 21* (2010), 440–447. 2

[CJMT08] COUR T., JORDAN C., MILTSAKAKI E., TASKAR B.: Movie and script: Alignment and parsing of video and text transcription. In *ECCV* (2008). 1

[Col02] COLLINS M.: Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Empirical Methods in Natural Language Processing (EMNLP)* (2002). 2

[CT07] COUR T., TASKAR B.: Video deconstruction: Revealing narrative structure through image and text alignment. In *NIPS Workshop on the Grammar of Vision: Probabilistic Grammar-Based Models for Visual Scene Understanding and Object Categorization* (2007). 1

[dSGCP16] DE SOUZA C. R., GAIDON A., CABON Y., PEÑA A. M. L.: Procedural generation of videos to train deep action recognition networks. *CoRR abs/1612.00881* (2016). URL: http://arxiv.org/abs/1612.00881. 4

[GRC15] GALVANE Q., RONFARD R., CHRISTIE M.: Comparing film-editing. In *Eurographics Workshop on Intelligent Cinematography and Editing, WICED '15* (Zurich, Switzerland, May 2015), Eurographics Association, pp. 5–12. URL: https://hal.inria.fr/hal-01160593, doi:10.2312/wiced.20151072. 3

[GRLC15] GALVANE Q., RONFARD R., LINO C., CHRISTIE M.: Continuity Editing for 3D Animation. In *AAAI Conference on Artificial Intelligence* (Austin, Texas, United States, Jan. 2015), AAAI Press, pp. 753–761. URL: https://hal.inria.fr/hal-01088561. 3

[GSSD09] GUPTA A., SRINIVASAN P., SHI J., DAVIS L. S.: Understanding videos, constructing plots - learning a visually grounded storyline model from annotated videos. In *CVPR* (2009). 1

[HO82] HOFFMANN C. M., O'DONNELL M. J.: Pattern matching in trees. *J. ACM 29*, 1 (1982), 68–95. doi:http://doi.acm.org/10.1145/322290.322295. 1

[JKS*15] JOHNSON J., KRISHNA R., STARK M., LI L.-J., SHAMMA D. A., BERNSTEIN M., FEI-FEI L.: Image retrieval using scene graphs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015). 2

[KM05] KELLAND M., MORRIS L.: *Machinima: Making Movies in 3D Virtual Environments*. Cambridge: The Ilex Press, 2005. 3

[LB93] LOYALL A., BATES J.: Real-time control of animated broad agents. In *In Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (1993). 4

[LB97] LOYALL A. B., BATES J.: Personality-rich believable agents that use language. In *Proceedings of the First International Conference on Autonomous Agents* (New York, NY, USA, 1997), AGENTS '97, ACM, pp. 106–113. URL: http://doi.acm.org/10.1145/267658.267681, doi:10.1145/267658.267681. 2

[LMSR08] LAPTEV I., MARSZALEK M., SCHMID C., ROZENFELD B.: Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition* (2008), pp. 1–8. doi:10.1109/CVPR.2008.4587756. 1

[Low05] LOWOOD H.: High-performance play: The making of machinima. In *Videogames and Art: Intersections and Interactions*, Clarke A., (eds.) G. M., (Eds.). Intellect Books (UK), 2005. 3

[Nit09] NITSCHE M.: *Video Game Spaces: Image, Play, and Structure in 3D Worlds*. MIT Press, 2009. 3, 4

[NM04] NITSCHE M., MAUREEN M.: Play it again sam: Film performance, virtual environments and game engines. In *Visions in Performance: The Impact of Digital Technologies*, Carver G., Beardon C., (Eds.). Swets & Zeitlinger, 2004. 4

[RGB15] RONFARD R., GANDHI V., BOIRON L.: The prose storyboard language: A tool for annotating and directing movies. *CoRR abs/1508.07593* (2015). URL: http://arxiv.org/abs/1508.07593. 2, 4

[Ron04] RONFARD R.: Reading Movies An Integrated DVD Player for Browsing Movies And Their Scripts. In *ACM Multimedia* (New York City, United States, 2004), ACM, (Ed.), ACM. URL: https://hal.inria.fr/inria-00545143. 2

[RRTS15] ROHRBACH A., ROHRBACH M., TANDON N., SCHIELE B.: A dataset for movie description. *CoRR abs/1501.02530* (2015). URL: http://arxiv.org/abs/1501.02530. 1

[RTT03] RONFARD R., TRAN-THUONG T.: A framework for aligning and indexing movies with their script. In *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)* (Baltimore, MD, United States, July 2003). URL: https://hal.inria.fr/inria-00423417. 2

[SLO07] SALWAY A., LEHANE B., O'CONNOR N. E.: Associating characters with events in films. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval* (New York, NY, USA, 2007), ACM, pp. 510–517. doi:http://doi.acm.org/10.1145/1282280.1282354. 1

[SVL14] SUTSKEVER I., VINYALS O., LE Q. V.: Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada* (2014), pp. 3104–3112. 4

[SWCS08] SHI Q., WANG L., CHENG L., SMOLA A.: Discriminative human action segmentation and recognition using semi-markov model. In *IEEE Conference on Computer Vision and Pattern Recognition* (June 2008), pp. 1–8. doi:10.1109/CVPR.2008.4587557. 2

[VRD*15] VENUGOPALAN S., ROHRBACH M., DONAHUE J., MOONEY R., DARRELL T., SAENKO K.: Sequence to sequence – video to text. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2015). 4

[WBR07] WEINLAND D., BOYER E., RONFARD R.: Action Recognition from Arbitrary Views using 3D Exemplars. In *ICCV 2007 - 11th IEEE International Conference on Computer Vision* (Rio de Janeiro, Brazil, Oct. 2007), IEEE, pp. 1–7. URL: https://hal.inria.fr/inria-00544741, doi:10.1109/ICCV.2007.4408849. 2

[YB08] YE P., BALDWIN T.: Towards automatic animated storyboarding. In *AAAI* (2008), pp. 578–583. 4