Toward Emergent Representations for Video

Ryan Shaw and Marc Davis

University of California at Berkeley School of Information Management and Systems

Garage Cinema Research

102 South Hall, Berkeley, CA 94720-4600, USA

http://garage.sims.berkeley.edu

{ryanshaw, marc}@sims.berkeley.edu

ABSTRACT

Advanced systems for finding, using, sharing, and remixing video require high-level representations of video content. A number of researchers have taken top-down, analytic approaches to the specification of representation structures for video. The resulting schemes, while showing the potential of high-level representations for aiding the retrieval and resequencing of video, have generally proved too complex for mainstream use. In this paper, we propose a bottom-up, emergent approach to developing video representation structures by examining retrieval requests and annotations made by a community of video remixers. Our initial research has found a useful degree of convergence between user-generated indexing terms and query terms, with the salient exception of descriptions of characters' corporeal characteristics.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems – *human factors, human information processing*; H.3.1. [**Information Storage and Retrieval**]: Content Analysis and Indexing – *abstracting methods, indexing methods, thesauruses*.

General Terms: Experimentation, Human Factors.

Keywords: Video retrieval, video annotation, video representation, remixing.

1. INTRODUCTION

This research is part of an effort to design a system to enable non-professional users to find, use, share, and remix video on the web. A key component in this system will be a shared metadata repository which holds pointers to fragments of video along with descriptive metadata for the content of those fragments. Users will both contribute descriptive metadata to the repository by annotating video content and make requests for video by providing descriptions of the desired content.

Such a system poses a couple of coordination problems. First there is the well-known problem of coordinating the vocabularies of metadata providers (users annotating video) with the vocabularies of video seekers [4]. For example, one user may describe a character in a shot as a "bad guy" while another may request shots portraying "villains." This problem may be solved in part by creating lists of related terms, either manually or by analyzing patterns of co-occurrence in textual descriptions.

More worrisome is the possibility that annotators will choose to describe attributes of content unrelated to the kinds of attributes video seekers specify in their requests. For example, an annotator may choose to spend all her time describing the lighting in a shot, neglecting to provide any information about the action it depicts. If a video seeker is interested in action, these annotations will be useless to him.

This could be solved by requiring annotators to annotate all possible facets of a given shot, but this solution makes the cost of voluntarily contributing to the metadata repository unacceptably high for a single annotator. A "divide and conquer" approach in which various facets are described by different annotators or groups of annotators could achieve the same goal, while keeping the granularity of individual contributions reasonably small. Such an approach requires representation structures which support the creation of compatible annotations by different users.

Another possible solution is to provide representation structures which can relate concepts used in annotations to those specified in queries. Rather than imposing an *a priori* representation structure on users, we have examined queries and annotations made by members of the target user community in order to determine what kinds of representation structures may be useful for ensuring coordination.

The goal of the experiment reported here was to determine the degree of divergence or convergence between the descriptions users use to annotate video content and the queries music video makers use to find content for remixing. The larger research this study is a part of aims to create systems that generalize and leverage the collective annotative efforts of enthusiast communities to enable the large scale annotation and retrieval of media on the web.

2. METHODOLOGY

The study was conducted through AMV.Org [1], a large (approximately 300,000 members) community of fans who appropriate Japanese animation (*anime*) content and re-edit it into music videos. (AMV is an acronym for *Anime* Music Video.) Editors use the site's discussion forums to make requests for desired content, which other fans then attempt to satisfy by suggesting suitable shots from various *anime* releases.

2.1 Request Analysis

We analyzed 220 requests for *anime* shots made on the AMV.Org "AMV Suggestions" discussion forum from July 2002 to April 2005 using a two-pass approach [5]. We first examined 100 requests and manually categorized them according to the attributes used in specifying the request. We developed the attribute set in parallel with our examination of the user requests, and based it on the actual requests generated by the user community. Next this coding scheme was used to categorize the full set of 220 requests.

2.2 Annotation Experiment

We created a simple media player application which enabled users to select temporal and spatial regions of two one-minute *anime* clips and add free-text annotations. The application was implemented using Macromedia Flash so that it could be accessed through a standard web browser. Annotations and their corresponding selected regions were stored in a relational database as they were created. Users could see a list of all the annotations they had created and could select annotations in the list to see the associated regions highlighted in the media player. Annotations could not be removed once they had been added.

The first one-minute clip was a scene from the English-dubbed USA release of *Cowboy Bebop: Knockin' on Heaven's Door* (2001), a popular *anime* film used as source content for 568 music videos listed on AMV.Org. This scene was selected because of its high action content, relative lack of dialogue, and the likelihood that it would be well-known to fans. The second clip was a scene from the English-dubbed USA release of *Perfect Blue* (1997), a somewhat less well-known film used as source content for 250 music videos listed on AMV.Org. This scene primarily featured a dialogue between two characters.

We invited 60 survey respondents—who had provided contact information and expressed interest in participating further in the project—to use the tool to annotate the two clips. Invitees were given no guidelines on what or how to annotate beyond basic instructions on how to use the annotation tool. The instructions simply asked users to "Add as many tags as you think are necessary to describe the scene."

3. RESULTS

3.1 Request Analysis

The results of the request analysis are shown in Table 1. Requests specified two different kinds of attributes on average, with a combination of action and object being the most common (14% of requests). The most commonly specified attributes overall were appearance-related, especially: actions depicted, specific objects, and specific physical attributes of bodies or faces. Specific names of characters, films, or TV series were less common than we expected.

Table 1. Shot request attributes and their frequencies.

	Attribute	Examples	Requests containing attribute	
			n	%
Appearance -related	action	breakdancing; committing suicide	89	40
	object	straitjacket; pancakes	64	29
	body/face	adult female; werewolves	59	27
	setting/era	a bar; in the darkness; a jungle	29	13
	number	two characters; large group	8	4

			Requests containing attribute	
Subject -related	plot/theme	one character dumping another; heroic characters working toward a common good	26	12
	mood/emotion	silly; psychedelic; sad	19	9
	role	villain; father; maniac	17	8
	genre	sci-fi; romance; *yaoi*	11	5
Specific names	cast member name	Vash; Tenchi	20	9
	film/series name	*Escaflowne; Trigun*	20	9
Production- related	camera/animation style/etc.	camera scrolls up; zoom in or out; loopable	9	4

3.2 Annotation Experiment

Over the two weeks of the study, 25 respondents created a total of 244 annotations for the two clips. Of these annotations, 210 (86%) specified particular temporal regions of the clip. In most of the cases where a specific region was not specified, it seems that this was due to user misunderstanding, since the annotation text seemed to refer to specific shots rather than the scene as a whole (although there were a few annotations which described the scene as a whole). Only 16 (7%) of the annotations specified spatial regions. This low number was probably due to the fact that the instructions did not make it sufficiently clear that it was possible to make spatial selections. Since spatial region selection, unlike temporal extent selection, is not a common feature of the media manipulation tools with which the annotators were familiar, it is likely this functionality was simply overlooked.

A wide variety of annotation styles were observed. Some annotators wrote long descriptions consisting of complete sentences:

> Nina's so stressed out and confused that she breaks her cup in her hands and isn't phased by it but meerly shows her state of mind by wondering if it was real. It shows that she's struggling with reality in a very dramatic and scary fashion...like a psycopath ^_^ (Then again, she IS psycotic...and the stalker....THE STALKER! AHH!!)

Other annotators used short, "tag" style annotations. One annotator in particular eschewed the use of spaces, creating tags like "StrutLowAngle."

The descriptions were coded using the set of attributes we developed from the shot request analysis (see Table 2). Descriptions specified two different kinds of attributes on average, with a combination of action and object being the most common (14% of descriptions). The most commonly specified attributes overall were appearance-related, especially actions depicted, specific objects, and specific physical attributes of bodies or faces. Specific names of characters were mentioned

quite often, though the names of the films from which the clips were taken were not.

Table 2. Shot description attributes and their frequencies.

	Attribute	Examples	Annotations containing attribute	
			n	**%**
Appearance -related	action	escape jump; wakes up	109	45
	object	teacup; truck	58	24
	body/face	looks confused; bloody hands	42	17
	setting/era	highway; bedroom	28	11
	number	*never used*	0	0
Subject -related	mood/emotion	creepy; strange	52	21
	plot/theme	Faye realizes she is tracking the wrong person; Mima begins questioning reality	29	12
	role	evil guy; mom	4	2
	symbolism	stigmata	3	1
	genre	action thriller	1	0
Specific names	cast member name	Vincent; Rumi	79	32
	film/series name	*Cowboy Bebop; Perfect Blue*	4	2
Production -related	camera/animation style/etc.	nice drawing; close-up	40	16
	audio/dialog	some quiet creepy music; "Is it real?" [quoting dialog]	15	6
	remix suitability	ripe for Photoshopping; made for AMV	4	2

Three new attributes were observed that had not been encountered in the shot requests. Three annotators described a shot focusing on a woman's bloody hands as "stigmata" or "Christ symbolism." This was considered sufficiently different from "plot/theme" to warrant a new attribute, "symbolism." A number of annotators also described the audio tracks of the clips, something not encountered in the shot requests (presumably since AMV editors do not generally use the audio tracks from *anime* content). Finally, one annotator described how certain clips might be used for creating AMVs, for example annotating spatial regions which (he claimed) could be easily altered using Adobe Photoshop.

4. IMPLICATIONS

The frequency distributions of attributes for the shot requests and the shot descriptions were quite similar (see Figure 1). As mentioned above, both requests and descriptions were most likely to specify actions depicted, with a combination of action and object being the most common type of annotation. This suggests that, for this community at least, severe coordination problems (such as the lighting vs. action case described in the introduction) will not be common.

However, some potential sources of coordination problems can be seen in the differences between the percentage of requests specifying physical attributes of bodies or faces (27%) and the number of descriptions specifying these attributes (17%), and between the percentage of requests specifying roles (8%) and the number of descriptions specifying these attributes (2%). (There is a similar gap for specific mentions of film or TV series names, but this is not a problem as this metadata can be assigned fairly easily through means other than manual annotation.)

These differences point to the need for a representation structure which supports linking the names of characters ("Vincent") to those characters' physical appearances ("tall," "male," "dark hair," "tattooed") and the roles they play ("villain"). This would enable the expansion of descriptions specifying names to also include these associated attributes. Since annotators seem to specify character names quite often, this approach could alleviate many potential coordination problems.

These results are quite encouraging, as they suggest that the semantic structure of this particular domain is such that there is a high degree of overlap between the ways users describe video content and the ways they query for it. Furthermore, the one notable area in which there is divergence between queries and descriptions can be addressed through the relatively simple technique of indexing physical appearances and roles by character names. Such an index could provide most of the value of a more complex ontology with far less labor required to build it. In fact, it seems quite likely that enthusiast communities like *anime* fans would enjoy creating such an index, just as they currently enjoy creating things like episode guides and character biographies.

In addition to the primary finding discussed above, we also found that a significant number of descriptions and requests specified moods, plots, or themes. This was surprising, as these are characteristics of video sequences that a remixer can change through the editing choices he or she makes. For example, it is not necessary to use source material from an *anime* series about a boy and a girl breaking up in order to create a music video about a boy and a girl breaking up—this narrative structure can be created whether or not it exists in the original content. Past representation structures for video [2, 3, 6] have been designed to clearly distinguish between the denotative and connotative meanings of video content, in order to make this kind of combinatory manipulation of sequence-dependent meaning possible. This tendency to specify moods, plots, and themes in descriptions and requests may reflect the fact that fans often seek to comment on original storylines known to their audiences. For example, a remixer could be creating a tribute to classic break-up stories, in which case footage is needed to refer to the plot or theme of the original narrative, not to create a new narrative. This suggests that a successful representation structure will allow users to describe and request subjective, connotative meanings of video sequences as well as inter-subjective, denotative meanings.
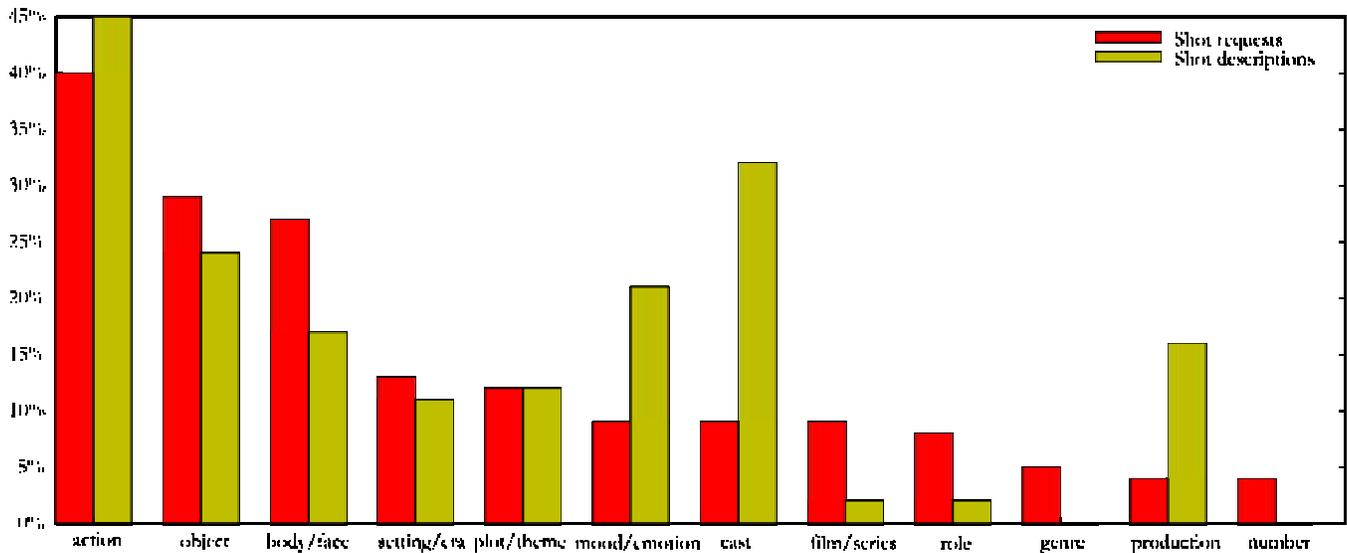
Figure 1. Shot request and shot description attributes and their frequencies.

5. RELATED WORK

A number of systems for representing video content in ways that support retrieval and remixing have been proposed. The earliest such system was developed by Bloch [2]. Bloch's representation framework includes the common *action* and *character* categories used by the AMV community, but does not make a clear distinction between characters and objects, nor does it provide a way to relate character names to appearances or roles.

Media Streams [3] uses an iconic visual language for representing video content. It makes the crucial distinction between *actor* (the physical appearance of a character) and *role* (how a certain character functions or is expected to function in the narrative), as well as providing a grammar for relating these descriptions to the names of characters. Later work on Media Streams distinguishes and enables combinations of the following distinctions: *actor* (named real person playing the character); *body* (the physical characteristics of the actor's body); *character* (the named character); *role* (the more general role of the character the actor plays); *character's body* (the physical characteristics of the character's body); and *costume* (the clothing worn by the character). These distinctions allow Media Streams to describe such complex combinations as "the male actor Dustin Hoffman playing a female actor playing a soap opera character in *Tootsie*." Since the Media Streams system is focused on sequence-independent representations of content, however, it does not support descriptions of high level actions or events such as "committing suicide," nor does it provide ways to describe the mood or plot of video sequences.

The representation structure of the AUTEUR system [6] makes a similar distinction between character's identities, appearances, and roles. Furthermore, it provides support for representing narrative and thematic knowledge of the type often described and requested in our study. However, these higher-level representations are geared toward the automated generation of narrative sequences, not necessarily the sort of description and retrieval uses we are concerned with in the current study. As discussed above, a generated sequence which tells a particular story or provokes a certain emotion may not meet the AMV editor's need, if she is counting on the audience's shared knowledge of the narrative or emotional content of the original work to provide the desired effect in the remixed work.

6. FUTURE WORK

The emergence of actual communities of practice engaged in the appropriation and reuse of video content on a large scale finally allows researchers to take an empirical approach to the development of representation structures for video. We hope to expand upon the initial investigations presented here by developing a system which will allow us develop representation structures by initially providing a community of users with a minimal structure, and then iteratively observing their use of it and modifying or extending it as needed. Eventually we hope to develop tools that will enable the community not only to create descriptions which conform to representation structures developed by researchers, but to participate directly in the ongoing development of those representation structures as well.

7. REFERENCES

[1] AnimeMusicVideos.Org, http://a-m-v.org.

[2] Bloch, G. R. From Concepts to Film Sequences. In *Proceedings of RIAO (RIAO '88)* (Cambridge, MA, March 21-22, 1988). 760-767.

[3] Davis, M. *"Media Streams: An Iconic Visual Language for Video Representation."* In *Readings in Human-Computer Interaction: Toward the Year 2000*, eds. R. Baecker, J. Grudin, W. Buxton, and S. Greenberg. 854-866. 2nd ed., San Francisco: Morgan Kaufmann Publishers, Inc., 1995.

[4] Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S.T. The Vocabulary Problem in Human-System Communication. *Communications of the ACM 30*, 11 (1987), 964-971.

[5] Hertzum, M. Requests for Information from a Film Archive: A Case Study of Multimedia Retrieval. *Journal of Documentation 59*, 2 (2003), 173-174.

[6] Nack, F. *AUTEUR: The Application of Video Semantics and Theme Representation for Automated Film Editing.* Ph.D. Thesis, Lancaster University, Lancaster, UK, 1996.