Modeling Intent for Home Video Repurposing

Radhakrishna S.V. Achanta
muvee Technologies

Wei-Qi Yan
Columbia University

Mohan S. Kankanhalli
National University of Singapore

Amateur home videos rarely convey intent effectively, primarily because of the limitations of conventional consumer-quality video cameras and the difficulties of video postprocessing. The authors describe a general approach for video-intent delivery based on offline cinematography and automated continuity editing concepts and demonstrate its use with four basic emotions: cheer, serenity, gloom, and excitement.

The term *video intent* represents an idea, theme, or message in a video clip expressed through the use of filmmaking principles. Unlike commercial movies that tell complex stories and let us relive emotions, home videos are shot mainly to record life events and have a limited capacity to express intent.

There are three primary reasons for this failure of home videos to capture the desired intent. First, home videos are often spontaneous recordings of family events, and, unlike carefully directed commercial movies, rarely involve preplanning. Moreover, amateur home video footage can't match commercial motion picture generation in terms of technique and quality. Second, home users must capture videos in uncontrolled settings (in a crowd, a moving vehicle, and so on) with limited facilities, resulting in artifacts due to camera-shake and poor lighting.[1] Lastly, home videos can't express their content coherently because camcorders for home use don't provide advanced audio- and video-mixing functionalities.[2]

IntentMaker provides home video editors with simple tools for enhancing their videos. The system uses automatic intent-delivery techniques conveying a range of evident emotions (cheer, serenity, gloom, and excitement) using principles of cinema grammar, aesthetics, and video analysis. IntentMaker works at the semantic level using video content as represented by low-level features. We focus on adding intents when they're not obvious in footage or enhancing them using film-grammar-based intent-delivery techniques.

Enhancing intent in home videos

Commercial movies generally have the means to express the basic human emotions—so much so that an entire set of grammatical rules for making movies has evolved. Viewers interpret these rules of grammar subconsciously to grasp the movie maker's intent. A given video clip has several intents, depending on the content at different points in the time line. Users can accentuate, modify, or even drastically change existing intents. Generally, they have three ways to do this:

- *Artifact removal*. Home users of handheld video cameras tend to shoot videos that are shaky, blurry, overexposed, or have excessive zooms in and out. Users can interactively detect and remove such artifacts during postprocessing.[1]
- *Intent delivery*. The second phase of video processing focuses on accentuating an intent that isn't convincingly conveyed in the raw footage, and is this article's focus.
- *Intent enhancement*. Users who want to exploit the footage's full potential[2] or test their creative skills can enhance the content using special effects and suitable audio–video mixing.[3]

Home video users can convey their video intents using software such as Adobe Premiere, Microsoft Windows MovieMaker, or Ulead Video Studio for postprocessing. Some software, such as muvee Autoproducer, even allows automatic video editing and summarizing according to a user-selected style, relieving the user from the burden of conventional nonlinear editing. Our research focuses on automating the imparting of intents to video. We do this by automating the application of the underlying cinema grammar techniques to the video. Our system automates the process (unlike nonlinear editors), but doesn't use a template-based approach (as in muvee's autoProducer).

Intent delivery techniques

Cinema grammar is the well-studied and doc-

1070-986X/06/$20.00 © 2006 IEEE

umented set of conventions and techniques that professional movie makers use to convey a film's story, theme, or meaning (see the "Literature Survey" sidebar). These rules are built on the interrelation between the human mind's conscious and subconscious responses to various combinations of the five major aesthetic fields in motion pictures: light, color, space, time, and motion. These techniques' methodical nature encourages us to fit them into formulas and use computers to automate basic movie making and imparting intent.

Light and color

In movies, lighting manipulations affect the perception of the properties of the environment (morning, noon, outdoors, indoors, and so on) observable in the scene, the psychological viewpoint (such as a gloomy or scary situation or a cheerful atmosphere) and the context for a certain event. In this context, *falloff* often describes the amount of brightness and contrast in a frame and the rate of change from bright to dark in a scene. Falloffs resulting from lighting arrangements significantly impact a scene's spatial, tactile, and temporal interpretation.

Humans best understand color in terms of its hue, saturation, and intensity components. It's thus natural that movie jargon deals with color in these terms. In film, saturation variations serve to shift emphasis from objects to characters. Higher saturation values make you observe, whereas lower saturation values make you *feel* too. Movie makers therefore increase or decrease color saturation depending on the degree of emphasis they place on characters with respect to the environment. In general, they consider color temperature, information to be conveyed, symbolism, and emotional attributes when deciding a scene's color content.

2D and 3D space

Edge-magnetism effects cause an object of interest's relative positioning within the frame (*framing*) to affect viewers' perception of space, congestion, and discomfort with regard to what's presented in the frame. For example, asymmetry can affect the scene's emphasis.

The perception of gravity and the resulting feelings of stability or insecurity depend on the angle of the base (horizon) on which things rest or occur in a scene. Frame asymmetry and the objects' relative positioning in the frame also induce varying interpretations in the viewer's mind.

Humans best understand color in terms of its hue, saturation, and intensity components. It's thus natural that movie jargon deals with color in these terms.

Depth and space creation and their subsequent manipulation is another aesthetic field. The use of wide angle, long shot, or normal lenses varies space perception and its effects.

Deblurring objects selectively (*fog filtering*)—indicating a shift in the camera's focus—can emphasize a person's or object's importance; similarly, zooming in to a smaller background object can increase the object's emphasis.

Time and motion

Time is mostly a subjective measure in cinema. Shot-tempo variation and proper shot transition effects can control time perception. Film makers often use vectors such as long shadows or birds returning to their nests to indicate objective time. In this context, vectors refer to directional cues as used in filmmaking parlance. Vector cues direct our attention subconsciously to what the film maker wants us to see.

Subjective time includes pace (the event's perceived speed), tempo, rate, and rhythm (flow within and among constituent segments such as shots, scenes, and sequences).

Motion categories include primary (object versus background), secondary (camera motion), and tertiary (movement and rhythm introduced by shot changes). Film makers often use slow motion to make an event look surreal.

Shot transitions—mainly cuts, dissolves, and fades—play a large role in deciding a movie sequence's pace and rhythm. Cuts—that is, instantaneous change from one shot to another—can convey changes among past, present, and future, or the simultaneity of two parallel events. A dissolve is a gradual transition from shot to shot with overlapping images from the two shots. Dissolves help maintain sequence flu-

Literature Survey

Previous work has used computers to automatically understand and manipulate video. Parkes, for example, describes a system that captures the content of instructional video sequences.[1] Auteur manipulates video to fit a thematic specification.[2] It takes a theme and a starshot as input, and builds the scene around them by contiguously placing preannotated "heaps" of shots. Similarly, Sack and Davis generate a video sequence according to a story plan from a database of preannotated video clips.[3] Baecker et al. present a similar system for designing and authoring structured and unstructured movies such as documentaries.[4] Other work uses a storyboard approach to develop multimedia presentations, including videos.

Garage Cinema envisages automating the production of home-made movies.[5] As Davis rightly points out, video content representation is the key to higher manipulation of video information.[6] This requires an understanding of syntax and semantics. In addition, video shot sequencing creates new semantics that might not be present in the individual shots and might supersede or contravene existing semantics.

Some attempts to represent video content and capture semantics in existing cinema use a computational media aesthetics framework from many aspects[7]—for example, beat,[8] tempo,[9] and rhythm.[10] These frameworks extract the related feature-based parameters from the video content and use corresponding algorithms to evaluate the aesthetics. Adams presents further work on understanding video through cinema grammar.[11]

Adams, Venkatesh, and Jain record the human experience of video editing and use these computable experiences to create a new video clip with certain media aesthetics.[12] Their aim is to help the average user build media artifacts that faithfully communicate intent while harnessing the chosen medium's full expressive powers. They provide two narrative templates—each with differing affect parameters—one emphasizing an emotive/intense response from the audience and the other seeking to maintain a higher level of clarity.

Video grammar generally reveals a video's computable elements. For example, people typically think of video rhythm as the organization of time, and video rhythm elements as accelerate (attack), decelerate (decay), metrical, and free.[8]

The main problem in video-intent delivery is formalizing the intent and computationally transferring intents to a video. Salway and Graham extract information about characters' emotions in films using a method based on a cognitive theory of emotions linking a character's emotional states to environmental events.[13] They classify emotions into 22 types, producing a list of emotion tokens (keywords) for each emotion type. Plutchik and Kellerman have established mathematical models to measure intents and have used cognitive algebra to measure them.[14]

Prince[15] and Zettl[16] describe the grammar of cinema aesthetics. Venkatesh and Dorai developed functional expressions to describe these cinematic rules.[17]

Sharff structures cinema performance rules according to eight basic models: separation, parallel action, slow disclosure, familiar image, moving camera, multi-angularity, master shot discipline, and orchestration.[18] He concludes that "the group of images so organized should generate more meaning than the sum of the information contained in each shot."

Barry and Davenport merge *subject sense* knowledge—common-sense knowledge stored in the open, common-sense database—and *formal sense* knowledge—knowledge gleaned from practiced videographers—to provide on-the-spot shot suggestions.[19] They aim to give the resulting raw footage the potential to be sculpted into an engaging narrative. If taken, shot suggestions retain their own metadata about the given shot.

Kennedy and Mercer list many possibilities available to cinematographers for mapping high-level concepts, such as mood, into decisions about which cinematic techniques to use.[20] They present a semiautomated planning system that helps animators present intentions via cinematographic techniques. The system

idity or can cushion the effect of conflicting vectors presented in the two shots. Fades, or gradual blackening of image frames, can indicate time passage or the end of one event and start of another.

Cuts can maintain a sequence's high-energy pace, whereas fades have the opposite effect.

Audio

Audio, especially music, is a highly integral part of a narrative.[3] Audio can improve a scene's intent delivery or act as a vector for calling attention to or signaling anticipation of an approaching event. A large part of the work in continuity editing aims to maintain sound continuity—in terms of volume, pitch, and context—to fit the storyline and the viewers' aesthetic tastes.[3]

Modeling intent delivery for home videos

We've identified several tools and a framework for imparting intents to video.

Intent delivery tools

We categorize intent delivery techniques into two classes:

- offline cinematography, which deals with frame-level changes; and

operates at the metalevel, focusing on animator intentions for each shot, and uses film grammar—including lighting, color, framing, and pacing—to enhance expressive power.

Our video-repurposing work aims to impart an intent to raw video footage using cinema grammar principles, steering clear of complicated semantic interpretations of input video. To the best of our knowledge, our approach, and the types of intents it currently imparts, have no precedence.

References

1. A.P. Parkes, "The Prototype Cloris System: Describing, Retrieval, and Discussing Video Stills and Sequences," *Information Processing and Management*, vol. 25, no. 2, 1998, pp. 171-186.
2. F. Nack and A. Parkes, "Auteur: The Creation of Humorous Scenes Using Automated Video Editing," *Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI) 95 Workshop on AI Entertainment and AI/Alife*, AAI/AI-ED, 1995, pp. 82-89.
3. W. Sack and M. Davis, "IDIC: Assembling Video Sequences from Story Plans and Content Annotations," *Proc. IEEE Int'l Conf. Multimedia Computing and Systems*, IEEE CS Press, 1994, pp. 30-36.
4. R. Baecker et al., "A Multimedia System for Authoring Motion Pictures," *Proc. ACM Multimedia*, ACM Press, 1996, pp. 31-42.
5. M. Davis, "Editing out Video Editing," *IEEE MultiMedia*, vol. 10, no. 2, Apr.–June 2003, pp. 54-64.
6. M. Davis, "Knowledge Representation for Video," *Proc. 12th Nat'l Conf. Artificial Intelligence* (AAAI 94), AAAI Press, 1994, pp. 120-127.
7. B. Truong, S. Venkatesh, and C. Dorai, "Application of Computational Media Aesthetics Methodology to Extracting Color Semantics in Film," *Proc. ACM Multimedia*, ACM Press, 2002, pp. 295-298.
8. B. Adams, S. Venkatesh, and C. Dorai, "Finding the Beat: An Analysis of the Rhythmic Elements of Motion Pictures," *Int'l J. Image and Graphics*, vol. 2, no. 2, 2002, pp. 215-245.
9. B. Adams, C. Dorai, and S. Venkatesh, "Formulating Film Tempo: The Computational Media Aesthetics Methodology in Practice," chap. 4, *Media Computing: Computational Media Aesthetics*, Springer, 2002.
10. B. Adams, C. Dorai, and S. Venkatesh, "Automated Film Rhythm Extraction for Scene Analysis," *Proc. Int'l Conf. Multimedia and Expo* (ICME), IEEE CS Press, 2001, pp. 1192-1195.
11. B. Adams, *Mapping the Semantic Landscape of Film: Computational Extraction of Indices through Film Grammar*, doctoral thesis, Curtin Univ. of Technology, Australia, 2003.
12. B. Adams, S. Venkatesh, and R. Jain, "IMCE: Integrated Media Creation Environment," *Proc. Int'l Conf. Multimedia and Expo* (ICME), IEEE CS Press, 2004, pp. 835-838.
13. A. Salway and M. Graham, "Extracting Information about Emotions in Films," *Proc. ACM Multimedia*, ACM Press, 2003, pp. 299-302.
14. R. Plutchik and H. Kellerman, *Emotion Theory, Research, and Experience: The Measurement of Emotions*, vol. 4, Academic Press, 1989.
15. S. Prince, *Movies and Meaning: An Introduction to Film*, Allyn and Bacon, 1997.
16. H. Zettl, *Sight, Sound, Motion: Applied Media Aesthetics*, 3rd ed., Wadsworth Publishing, 1999.
17. D.Q. Phung, S. Venkatesh, and C. Dorai, "Hierarchical Topical Segmentation in Instructional Films Based on Cinematic Expressive Functions," *Proc. ACM Multimedia*, ACM Press, 2003, pp. 287-290.
18. S. Sharff, *The Elements of Cinema (Toward a Theory of Cinesthetic Impact)*, Columbia Univ. Press, 1982.
19. B. Barry and G. Davenport, "Documenting Life: Videography and Common Sense," *Proc. Int'l Conf. Multimedia and Expo* (ICME), IEEE Press, 2003.
20. K. Kennedy and R.E. Mercer, "Using Cinematography Knowledge to Communicate Animator Intents," *Proc. 1st Int'l Symp. Smart Graphics*, ACM Press, 2001; http://www.dfki.de/~krueger/sg2001/schedule/Kennedy.pdf.

- automated continuity editing, which copes with shot-based changes and audio use.

Offline cinematography. In cinema parlance, *cinematography* describes a frame's composition (including the objects in the frame and their relative positions), the dominant hues and their saturation and intensity levels, lighting angles, and camera motion.

Offline cinematography is the set of intent-delivery techniques that lets users alter several compositional units during postprocessing (hence the term offline). We've identified a subset of possible offline cinematography techniques based on our study of cinema grammar.

Our lighting manipulation tools include the following:

- *Brightness change.* For brightness change, we represent an image by f(x, y), with $x = 1, 2, \ldots, W$ and $y = 1, 2, \ldots, H$ denoting the position of pixels in the image, where W and H are the video frames' width and height. To get a brighter image $b(x, y)$, we multiply the image $f(x, y)$ by a constant $c \in \lfloor 0, 1 \rfloor$, and vice versa for a darker image: $b(x, y) = c \times f(x, y) + (1.0 - c) \times 255; c > 0.5$.

- *Contrast change.* We obtain contrast change by

Because amateur film makers don't use multiple cameras or vector cues, our intent-delivery work attempts to automate a small subset of continuity editing-based processing techniques.

making pixels above a threshold T brighter, and the rest darker than their existing values. We define the contrast Con by the relevant luminance using $Con = L_{max} - L_{min}/L_{max} + L_{min,}$ where L_{max} and L_{min} are maximum and minimum luminance. We obtain the luminance Lum using the equation, $Lum = 0.2125 \times R + 0.7154 \times G + 0.0721 \times B$, where (R, G, B) is the red, green, and blue image pixel color.

- *Variable lighting*. We can vary lighting both across a frame and across a shot using a random function $v(t) \in [0, 1]$, varying according to y in place of a constant c, where t is the time: $b(x, y) = v(t) \times f(x, y) + (1.0 - v(t)) \times 25$; $v(t) > 0.5$.

The color manipulation tools let us convert the RGB space to HSI space to bring about changes with respect to the color's hue, saturation, and intensity.

Our space manipulation tools include the following:

- *Image zooming*. We simulate the camera zooming in and out using an interpolation technique.
- *Image tilting/rotation* (for edge effects/magnetism). We perform these actions using any of the standard image rotation algorithms. We follow tilting by image enlargement to cull undesired portions of the video frame.
- *Selective image cropping* (for frame placement). After performing a simple image enlargement, we crop the desired region.
- *Video flipping*. Flipping each image of a video generates a different viewpoint for the video.

We created Table 1 using these movie-making principles. The processing options listed for each intent conform to the movie-making principles.

The operations mainly reflect the operations for achieving the target emotions. We ignore some minor and less useful operations.

We classify the operations performed on the input video into operations for offline cinematography and those for automated continuity editing. Our algorithms (presented later)

Table 1. Intent delivery tools for selected emotions.

Operation	Emotion			
	Cheer	Serenity	Gloom	Excitement
Offline cinematography-related operations				
Brightness change	Increase	Decrease	Decrease	Increase
Contrast change	Increase	Decrease	Decrease	Increase
Saturation change	Increase	Decrease	Decrease	Increase
Intensity change	Increase	Decrease	Decrease	Increase
Frame blurring	None	Increase	Increase	None
Zooming in and out	Add zoom-in	None	None	Add zoom-in and zoom-out
Frame rotation	Add	None	Add	None
Frame flipping	None	Increase	Increase	None
Continuity editing-related operations				
Video tempo	Speed up	Slow down	Slow down	No change
Shot reordering	Yes	No	Yes	Yes
Transition frequency	More	Less	Less	More

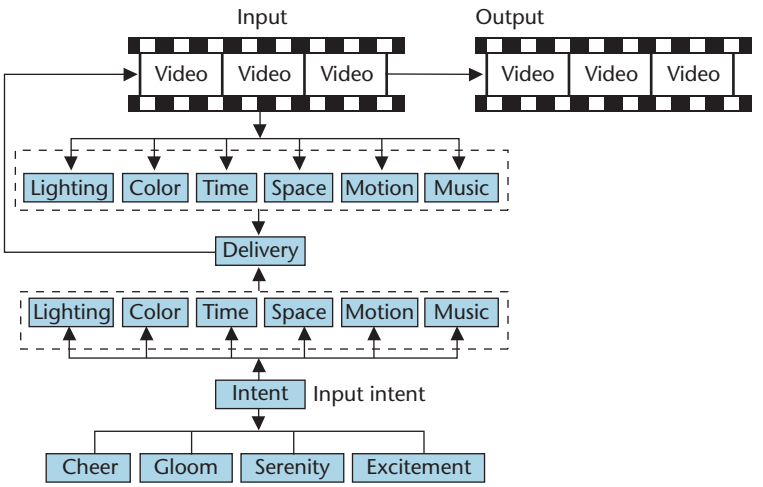

Figure 1. Intent delivery framework.

demonstrate our application of the operations. The degree to which we apply them depends on the video properties obtained from the video analysis. For example, for excitement intent, we zoom-in on low-action-frames (LAFs), rotating them to increase their action. We can also speed up the video in the vicinity of these frames by dropping them.

Music plays an important part in maintaining continuity and imparting intent. We therefore choose appropriate music clips to accompany the processed video. In this work we choose the appropriate music; however, our earlier work provides a starting point for automating this step.[3]

Automated continuity editing. Continuity editing is the method of composing the entire movie with several constituent shots without introducing any noticeable jarring disturbances, so as to narrate the story smoothly. Continuity-editing techniques let centuries pass in a few seconds or suspense build with appropriate shot tempo and duration (among other things).

Continuity editing has standard methodologies. Much of continuity editing requires shooting raw footage with multiple cameras to capture different parts of the scene with different views, as well as a complex use of various vector cues. Film makers must also include several important aspects regarding graphical and index vector cues (as well as lighting and color choices) at the time of shooting.

Because amateur film makers don't use multiple cameras or vector cues, our intent-delivery work attempts to automate a small subset of continuity editing-based processing techniques, including shot cuts, shot reordering, fade in and out, and wipes. For each example intent, IntentMaker chooses a shot tempo of appropriate length and accompanying transitions or cuts to maintain continuity and correct delivery. This ordering of shots belonging to the same context can also be nonsequential. In this case, the continuity-maintenance factor is mainly the accompanying music.

We use intent-delivery tools to convey the intents of the four basic emotions (cheer, serenity, gloom, and excitement) to a given piece of raw video footage. Home videos don't have to follow a story line, and different parts of the raw footage record different events that might convey any specific intent. IntentMaker aims only to impart or enhance intent in raw footage, in part or in whole, as the user chooses. For example, the user might not want to impart the gloom intent to the entire video, but only to those parts containing, for example, a funeral or the aftermath of a natural disaster.

Framework for automatic intent delivery

Figure 1 illustrates the intent-delivery framework. The user inputs a video clip to be placed within a specific story line. The user specifies the intent he or she wants the clip to express from the four choices (cheer, serenity, gloom, and excitement). The framework preprocesses the clip to extract certain low-level features (brightness, degree of color spread, amount of motion in frames and action regions, and so on) used in deciding the video-processing options and the parameters for regulating them. Currently, the focus is on processing a single clip only, implying no change of context.

The framework uses the extracted features for

Input : Captured video
Output : Extreme action regions and frames

Procedure:
Step 1. Find image difference between successive frames.
Step 2. Obtain the binary thresholded image.
(Camera motion is present if the number of bright pixels in the thresholded image is greater than the total number of pixels in the frame by another threshold.)
Step 3. Pass a sliding window one-third of the image's size over the entire image. The maximum activity is in the image subregions with the highest sum of pixel values.
Step 4. Average the coordinates of the image region over the number of frames on which the desired image transformation is to be applied.

Figure 2. Algorithm 1 finds extreme value action regions or frames.

Table 2. Intent-delivery symbols for home videos.

Symbol	Description
N	Number of frames in the clip
N	Number of a certain frame
$[f_s, f_e]$	Frame range from start value f_s to end value f_e
S_o	Original saturation value
S_n	New saturation value
I_o	Original intensity value
I_n	New intensity value
HAF	High-action frame
LAF	Low-action frame
T_1, T_2, T_3	Threshold
r, g, b	Red, green, and blue (RGB) values of each pixel
h, s, i	Hue, saturation, and intensity (HSI) values of each pixel
cf	Fixed set of consecutive frames
Step	Multiple used for linearly varying a quantity

Table 3. Intent-delivery functions in home videos.

Function	Description
RGB2HSI(r, g, b)	Color space conversion for the RGB color space
HSI2RGB(h, s, i)	Color space conversion for the HSI color space
SaturationPerFrame(f_n)	Return average saturation per frame
IntensityPerFrame(f_n)	Return average intensity per frame
ContrastChange(degree)	Change contrast according to the degree
EnlargeFrame[$R[f_s, f_e]$]	Enlarge frames with the same scale
Zoom[$R[f_s, f_e]$]	Performs zoom-in, then zoom-out
Blur(degree, f_s, f_e)	Blurs frame by the input degree
Rotate(f_n, α)	Tilts frame f_n by the angle α required
FadeIn($R[f_s, f_e]$)	Fade into the scene within frame range
FadeOut($R[f_s, f_e]$)	Fade out of the scene within frame range
Wipe($R[f_s, f_e]$)	Creates a wipe within given frame range
CrossFade($R[f_s, f_e]$)	Creates cross-fade shot transition
CreateFrame(f_n)	Creates new frame in the vicinity of f_n
DropFrame(f_n)	Deletes new frame in the vicinity of f_n

Note: We use the function R[·]to execute a function within a range $[f_s, f_e]$.

offline cinematography and automated continuity editing in such a way as to generate a processed video that conforms to the rules of cinema grammar, and outputs a video clip of the desired intent. It adds a guideline music clip to the final output. These techniques currently rely on a few simple features:

- average color saturation per frame to decide color changes,
- average luminous intensity per frame to decide lighting changes, and
- frame differencing to decide degree of motion.

An important aspect of the automation is finding each frame-processing tool's operating constraints. For example, IntentMaker shouldn't perform a tilt operation on a frame if it causes the image's most important part to be left out of the frame bounds. This requires finding the bounds of the video sequence's most important features, which we determine by identifying high-action regions located in binary frame difference images, or detecting skin regions using a Gaussian skin classifier. Algorithm 1 (see Figure 2) helps us find the extreme value (highest or lowest) action regions or frames.

Intent generation

IntentMaker uses a small subset of techniques mentioned in cinema grammar literature. The video content must be suited to the particular emotional intent, although interesting results are sometimes possible otherwise. Table 2 lists the symbols used in video-intent delivery.

We perform some preliminary analysis to calculate motion activity (based on simple frame differencing) and color histograms. This helps find LAFs and high-action frames (HAFs) in the raw footage. Table 3 describes the video-processing functions that we use in our algorithms.

We compute video-intent delivery of the four basic emotions using the linkages in Table 1. We perform this task according to our definitions of the intent contributions of video features, video-intent delivery tools based on video-feature analysis, and video-intent delivery functions and their variants (Tables 2 and 3). Video-intent delivery tools help us map the four intents to the four algorithms.

- *Cheer*. We use Algorithm 2 (see Figure 3) to

generate cheer, which is typically a sense of joviality reflected in video frames. Color saturation and high brightness levels make a given video scene look cheerful. A higher pace and tempo add to this feeling. The saturation and intensity values increase as we associate higher values with positive situations. Cheerful situations are dynamic, so we introduce shot transitions and drop frames in the vicinity of LAFs to minimize passiveness.

- *Serenity*. The serenity intent conveys peace and calm in a scene. Such scenes are low action, and color saturation is generally lower to keep the focus on the subject and context rather than the surroundings and other events. Serenity can sometimes resemble gloom. Using the right type of audio and introducing some surrealism with Algorithm 3 (see Figure 4) can help clarify the difference. Blurring at the clip's beginning and end adds some surrealism, conveying a feeling of peace or the passage of a long period of time. Wipes and cross-fades convey that the context persists despite prolonged viewing.

- *Gloom*. A gloomy scene conveys sadness, translating feature-wise as low scene energy (and thereby less enthusiasm associated subconsciously with the scene). Color saturation is reduced, brightness is lessened, and action progressively slowed. We also introduce slowness and constancy in the scene by ensuring that the energies continue to decrease as the scene progresses, and by adding frames in the vicinity $0.1 \times cf$ frames before and after the HAFs. A fade-in at the beginning and a fade-out at the end accentuates the darkness associated with the scene, and hence the gloom associated with the video clip. Algorithm 4 (see Figure 5) illustrates this process.

- *Excitement*. Excitement is a positive emotion, conveyed cinematically with high-motion content and upbeat music. High-energy scenes full of events require a lot of the viewer's attention and generate excitement. We impart this sort of high energy using tilts and zooms, high contrasts, and bright colors. We perform frame-tilting operations at LAFs so that gravity effects induce interpretation of high action in the frame, thereby increasing excitement in them. In both LAFs and HAFs, zooms focus on the high action in the frame, enhancing the excitement from the additional motion, as

```
Input  : Captured video
Output : Repurposed video with enhanced CHEER intent

Procedure:
for 0 : N do
    RGB2HSI(r, g, b);
    S_n = (1 + v_s) * S_0; v_s ∈ [0, 1];
    I_n = (1 + v_i) * I_0; v_i ∈ [0, 1];
    HSI2RGB(h, S_n, I_n);
    if HAF then
        FlipVideo(R[f_s, f_e); Enlarge(R[f_s, f_e);
    end

    if LAF then
        Dropframe(f_n);
    end
end
```

Figure 3. Algorithm 2 delivers the cheer emotion for home videos.

```
Input  : Captured video
Output : Repurposed video with enhanced SERENITY intent

Procedure:
Blur(degree - -; [0, c_f + 10]);
for 0 : N do
    if AvgAction > T_a then Wipe(R[f_s, f_e]);
    else CrossFade(R[f_s, f_e]);
    if HAF then
        CreateFrame(f_n);
    end
end

Blur(degree + +; [N - cf, N]);
```

Figure 4. Algorithm 3 delivers the serenity emotion for home videos.

```
Input  : Captured video
Output : Repurposed video with enhanced GLOOM intent

Procedure:
FadeIn(R[0, cf]);
Blur(degree + +; [0, cf + 10]);
for 0 : N do
    RGB2HSI(r, g, b);
    if SaturationPerFrame() > T_s then
        S_n = (1 - v_s * step + +) * S_0;
        v_s Œ [0, 1];
    end
    if IntensityPerFrame() > T_i then
        I_n = (1 - v_i * step + +) * I_0;
        v_i Œ [0, 1];
    end
    HSI2RGB(h, S_n, I_n);
    if HAF then
        CreateFrame(f_n);
    end
end

FadeOut(R[N - cf, N)
```

Figure 5. Algorithm 4 delivers the gloom emotion for home videos.

Input: Captured video
Output: Repurposed video with enhanced EXCITEMENT intent

Procedure:
for 0 : *N* **do**
  RGB2HSI(r, g, b);
  $S_n = (1 + v_s) * S_0$; $v_s \in [0, 1]$;
  $I_n = (1 + v_i) * I_0$; $v_i \in [0, 1]$;
  HSI2RGB(h, S_n, I_n);
  if LAF **then**
    Rotate(f_n, q);
  end
  if HAF **then**
    Zoom($R[(f_n - 30), (fn + 30])$);
  end
end

Figure 6. Algorithm 5 delivers the excitement emotion for home videos.

Figure 7. "Warrior" video: (a) original and (b) processed with cheer intent.

Figure 8. "Nadia" video: (a) original and (b) processed with excitement intent.

Algorithm 5 (see Figure 6) shows. In our video-intent delivery system, contrast changes introduce higher energy into the scene.

Video-intent delivery results

Figures 7 through 10 present frames from the processed video clips. In each figure there are two rows. The top row shows raw video frames, while the bottom row shows the same frames after imparting the respective intent-based-processing on to it. Among the four clips shown for each video, the first two from the left in each row are LAFs and the next two are HAFs. For an active demonstration of the results, you can download the video clips from http://www.comp.nus.edu.sg/~mohan/intent/index.htm.

Figure 7b illustrates one of the system's drawbacks. Parts of the person's face are out of the frame's range, because no mechanism is in place to detect a face and place a constraint on the degree and area of enlargement, rotation, and so on.

We processed the clips illustrated in the figures with the intent that was rated as best according to the survey results shown in Table 1. As Figure 8b shows, the saturation and intensity values are higher for the processed clip than for the original.

Figure 9 shows a truck passing a wagon. We classified this frame as high action, and the algorithm tried to emphasize this part by enlarging the area. However, this part of the video is semantically less important.

In Figure 9, some corresponding frames appear to have different content. Frame dropping and operations such as crossfading and reducing the net number of frames can cause this effect. In Figure 10, the gradual change in intensity and saturation is evident.

To evaluate our results, we asked a set of 10 users to rate the four sets of processed video clips on their ability to express intent. We first showed the participants the processed clip and asked them to identify the clip's intent. We recorded the degree to which the user agreed with the actual intent as a score between 1 and 10. We took two scores—one with music and one without. Table 4 presents the average of the two scores.

The survey subjects generally agreed that the processed video conveyed intent more effectively than the original video, even without music. The presence of music, however, strongly reinforces the intent and removes ambiguities (say, between serenity and gloom intents). The results also show that force-fitting an intent onto a video whose content is not suited for it (for example, using the cheer intent for the "Walk" video clip or the gloom intent for the "Nadia" clip) produces poorer results.

From Table 4, we also see that the ratings basically reveal the emotions of a given video, and also correctly reflect the results of the intent-delivery scheme. This table is useful for compar-

ing the computing results with human perception. Using the video-intent delivery algorithms, we can attempt to manipulate video semantics by changing video features in conformance with cinema grammar rules. Thus the statistics in Table 4 reveal the linkage between human feelings and the algorithms.

Future work

Our future work will look at increasing the range of available high-level intents and the number of low-level image and video-processing tools. Video preprocessing can involve low-level features such as robust object detection, tracking, and segmentation algorithms to significantly improve intent delivery. **MM**

Figure 9. "Wagon" video: (a) original and (b) processed with serenity intent.

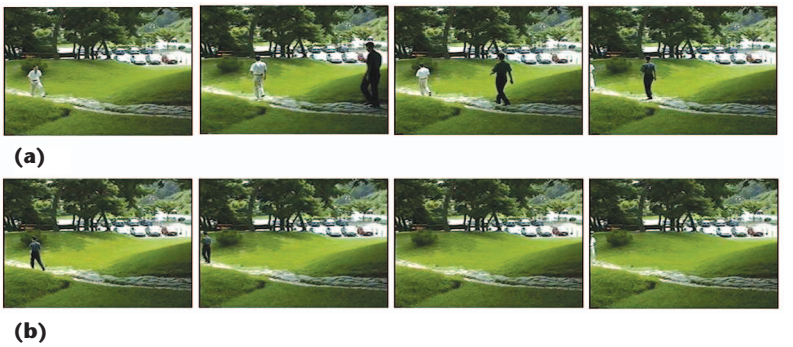

Figure 10. "Walk" video: (a) original and (b) processed with gloom intent.

Table 4. Video's ability to convey intent, rated from 1 to 10.

Clip	Intent				Maximum	Minimum
	Cheer	Gloom	Serene	Excite		
Nadia	6.3	6.0	6.3	8.1	8.1	6.0
Wagon	7.7	6.7	7.2	7.2	7.7	6.7
Walk	5.8	7.6	6.7	7.4	7.6	5.8
Warrior	7.7	7.3	7.4	6.9	7.7	6.9
Maximum	7.7	7.6	7.4	8.1	8.1	–
Minimum	5.8	6.0	6.3	6.9	–	5.8

Acknowledgments

We deeply appreciate the anonymous reviewers' many constructive suggestions.

References

1. W.Q. Yan and M.S. Kankanhalli, "Detection and Removal of Lighting and Shaking Artifacts in Home Videos," *Proc. ACM Multimedia*, ACM Press, 2002, pp. 107-116.
2. C. Dorai and S. Venkatesh, "Computational Media Aesthetics: Finding Meaning Beautiful," *IEEE MultiMedia,* vol. 8, no. 4, Oct.–Dec. 2001, pp. 10-12.
3. P. Mulhem et al., "Pivot Vector Space Approach for Audio-Video Mixing," *IEEE MultiMedia*, vol. 10, no. 2, Apr.–June 2003, pp. 28-40.

Radhakrishna S.V. Achanta is a software engineer at muvee Technologies, Singapore. His research interests are image and video processing and content-based retrieval. Achanta has a master's degree in computer science from the National University of Singapore.

Wei-Qi Yan is a postdoctoral research scientist at Columbia University. His research interests are multimedia systems and media security. Yan has a PhD in computer engineering from Academia Sinica.

Mohan S. Kankanhalli is a professor in the Department of Computer Science at the National University of Singapore. His research interests are multimedia information systems and information security. Kankanhalli has a PhD in computer and systems engineering from the Rensselaer Polytechnic Institute.

Readers may contact Wei-Qi Yan at wy2124@columbia.edu