Evolution-based Virtual Content Insertion

Ming-Che Chiang¹, Chia-Hu Chang², and Ja-Ling Wu¹,²

¹Department of Computer Science and Information Engineering
²Graduate Institute of Networking and Multimedia
National Taiwan University
Taipei 10617, Taiwan, R.O.C.
{blossom, chchang, wjl}@cmlab.csie.ntu.edu.tw

ABSTRACT
This demonstration presents an innovative framework for virtual content insertion in an interactive way. The virtual content is inserted into video with evolved animations according to predefined behaviors, and generates interactions with video contents. In order to reduce the intrusiveness and improve the impression for viewers, the evolution process is divided into distinct yet dependent phases, in which the virtual content evolves its appearances and behaviors according to the incremental interactions. Therefore, the augmented videos generated by the proposed system establish visually relevant connection between the inserted virtual content and the source videos, and increase the acceptability and the attractiveness at the same time. Moreover, the proposed system enables video owners to create entertaining personalized videos effectively, and engages viewers with the storyline.

Categories and Subject Descriptors
I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Object recognition, Shape; H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Video

General Terms
Algorithms, Experimentation, Human Factors

Keywords
Virtual content insertion, animation, simulated evolution, interaction

1. INTRODUCTION
With the dramatically boosting of online video service, the characteristics of easy to create and share videos have created phenomenal opportunities for advertisers and content owners seeking to monetize and personalize their video assets. By using the virtual content insertion technology, some existing online video sites provide new advertising services to increase the additional revenue with regard to advertising for online video services. However, it is observed that most related work and online services rarely involved how to represent the virtual content. It has been pointed out [1] that the inserted virtual content without the connection to the source video in terms of the semantic meaning or the interaction would bring about negative effects on audience’s viewing experience.

We demonstrate a novel system which constructs a visually relevant connection between the inserted virtual content and the source videos. The system vividly evolves the inserted virtual contents along an incremental and interactive evolution process, and provides users a creative way to enrich and enhance source video contents. Based on the concept of evolutionary biology, we assign characteristics about appearance and behaviors to the virtual content and animate it by interacting with the video content. The videos are considered as not only carriers of message conveyed by the virtual content but also the environment in which the lifelike virtual contents live. The interactions between the virtual content and the source video would produce an entertaining storyline accordingly. In our system, the evolution path is broken up into three distinct yet dependent phases, which are the cell phase, the microbe phase, and the creature phase.

In the beginning, the virtual content would be dropped into videos in a less-intrusive manner because of its simple appearance and behavior which are like a single-celled organism. Through the cell phase, the cell-like virtual content is subject to the background influence and acts passively. With the sense of vision being matured, the virtual content in the microbe phase would begin to discover and develop the synonymous ability to enhance its own colors and textures. In the creature phase, with the sense of hearing being matured, the virtual content would dance with the perceived aural stimuli, and interact with the moving salient object in an intelligent manner.

2. SYSTEM OVERVIEW
Figure 1 illustrates the framework of the proposed system, which consists of three stages: (1) the video content analysis stage for constructing multilayer feature space of each frame of the input video; (2) the virtual content analysis stage for assigning various appearance and behaviors in each phases based on the predefined evolution; (3) the animation generation stage for animating the moving path of the virtual content on the virtual layer.

Given the input video, the motion vectors are extracted and the video frames are segmented according to the colors and textures in [2]. With the frame profiling module, we construct the feature map of background, which describes that the extents of the influence in each segmentation region
for the virtual content are different. On the other hand, for the purpose of determining which object interacts with the virtual content, a region of interest (ROI) is localized in each video frame by combining various visual features. In addition to the visual information, aural saliency analysis is performed to detect the time unit with salient strength of the sound [3], so the virtual content would respond to the audio. For the selected virtual content, its appearance is analyzed for visually evolving in terms of shapes, colors and textures. Therefore, the virtual content would evolve into complete pattern gradually. The mass is defined as the area of the virtual content that has grown.

In the behavior modeling module, the virtual content is assigned life-like behaviors through the evolution process. In the cell phase, we make the virtual content a single-celled organism liked. The virtual content would perceive the force from the environment and evolve. To reduce the intrusive-ness, the color of virtual content is further harmonized with the background by the method proposed in [4], and simplified to single color by averaging. Moreover, the technique of morphing is used to generate realistic transitions during the evolution. In the microbe phase, the virtual content is assigned the ability to seek for salient object. The original colors of cell would be presented with the fading-in effect and the opacity would be modulated by a Gaussian function. At the same time, the salient object would be visually erased. In this way, the virtual content is simulated to obtain the color and texture by absorbing the energy of the salient object, just like the microbe organism. In the creature phase, the virtual content is simulated to own the ability to dance with the music or represent astonished expression while perceiving loud sound. Besides, the virtual content interacts with the moving salient object in an intelligent manner. The virtual content either tends to imitate the behavior of the moving salient object, or moves to the salient object for further interaction.

With the feature space of the input video frames and the life of the virtual content, our system automatically generates impressive animations with an evolution way on a virtual layer. Finally, the virtual layer, in which the virtual content is animated on, is integrated with the video layer.

3. EXPERIMENTS

We recorded a personal video, which contained only a person doing some actions. With the virtual content embedded by our system, the person can interactively play with the virtual content, as shown in Figure 2. The interaction seems to be an interactive game through the webcam. The occurred events depend on the properties of videos and the predefined genotype of the inserted virtual content. Therefore, a new storyline, which is induced by a series of interactive events, can be generated in the augmented video.

4. CONCLUSIONS

We have presented a novel system which provides a lifelike animation for inserting virtual content into general videos in a vivid way to enhance the impression and acceptability of virtual content. The evaluation results show that the produced videos improve the audience’s viewing experience to the original video content and engage viewers with interesting interactions.

The application scenarios of our system include the following. First, humans may interact with the virtual content via the webcam. Second, the system provides a platform for the users to create and visualize the script of personal stories. Third, it’s a tool to build unique personal videos with originality. Finally, the system makes virtual product placement more vivid with the interaction. We believe that the proposed innovative system for virtual content insertion would inspire users to create entertaining storylines and turn their videos into visually appealing ones.

5. REFERENCES

Evolving Virtual Contents with Interactions in Videos

Chia-Hu Chang\(^1\), Ming-Che Chiang\(^1\), and Ja-Ling Wu\(^{1,2}\)
\(^1\)Department of Computer Science and Information Engineering
\(^2\)Graduate Institute of Networking and Multimedia
National Taiwan University
Taipei 10617, Taiwan, R.O.C.
{chchang, blossom, wjl}@cmlab.csie.ntu.edu.tw

ABSTRACT
With the development of multimedia analysis, virtual content insertion has been widely used and studied for the video enrichment and advertising. However, how to less-intrusively insert a virtual content into general videos with an attractive representation is a significant and challenging problem in the field of virtual content insertion. In this paper, we present a novel virtual content insertion system which inserts virtual contents into videos with evolved animations according to predefined behaviors based on the concept of evolutionary biology. The videos are considered as not only carriers of message conveyed by the virtual content but also the environment in which the lifelike virtual contents live. Thus, the inserted virtual content interacts with video contents and triggers the artificial evolution. The evolution process is divided into distinct yet dependent phases, in which the virtual content evolves its appearances and behaviors with the incremental interactions. In this way, the proposed system constructs a visually relevant connection between the inserted virtual content and the source videos to reduce the intrusiveness and increase the acceptability and the attractiveness simultaneously. User studies show that the augmented videos produced by the proposed system improve the audience’s viewing experience to the original video content and engage viewers with the entertaining storyline.

Categories and Subject Descriptors
I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Object recognition, Shape; H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Video

General Terms
Algorithms, Experimentation, Human Factors

Keywords
Interaction, animation, simulated evolution

1. INTRODUCTION
With the technical advances in video coding and the rapid development of broadband network delivery, online video services are dramatically boosted. There are more and more users uploading various videos of their own with the uniqueness, creativity and interest to the video sharing sites, such as Youtube [1], to broadcast themselves and share the life experience with online audience in the Internet. The characteristics of easy to create and share videos have led a huge amount of video content distribution and have created phenomenal opportunities for advertisers and content owners seeking to monetize and personalize their video assets. Therefore, how to attract more online video audience and create additional revenue from the existing online video inventory is an emerging problem.

A few existing online video services start to develop video analytics and video editing tools for users to enrich and enhance the message conveyed in videos by inserting additional contents. For example, the Youtube analytics tool Hot Spots compares each video to other videos of similar length on YouTube to determine which points in a video are “hot”. On the other hand, Youtube provides video annotations for users to add speech bubbles, notes and spotlights overlays on videos. In addition, Google develops new video advertising called InVideo ad to try to get economic benefits from its video property. Generally, all the above mentioned additional contents which are inserted into the videos virtually can be considered as the virtual content compared to original content. According to different purposes, such as advertising, entertainment, and information enhancement, the inserted virtual contents could be the brand, commercial logos, interesting images, informative windows, or whatever message which content creators desired to broadcast. The more attractive a video is the more advertising revenue it can generate. Therefore, the virtual content insertion have received tremendous attention from both academia and industry sides.

By using the virtual content insertion technology, the existing online video sites, such as Overlay.TV [2] and Innovid [3], improve the viewing experience to the original content and provide new advertising services to increase the additional revenue with regard to advertising for online video services. In addition to the above mentioned online video services, efficient methods and automatic systems [4-6] for virtual content insertion have been studied in the past years. With the techniques of content-based video analysis, the geometrical relationships between the camera and the surface of the flat area in which the virtual content is projected

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
IMCE’09, October 23, 2009, Beijing, China.
Copyright 2009 ACM 978-1-60558-758-5/09/10 ...$10.00.
onto can be easily calculated. Based on the visual attention model, a virtual content insertion system was proposed in [7] to support various types of videos. In [8], the ViSA system was proposed to provide an advertising mechanism for tennis videos by taking psychology, computational aesthetics, and advertising theory into account. The representation of virtual content induces an ingenious interaction between the inserted virtual content and the moving player. In the existing work, static representation is widely selected for virtual contents insertion to reduce the induced intrusiveness. As a result, in order to grab the viewer’s attention, the insertion time and place are limited to some estimated points.

Obviously, how to less-intrusively insert the contextually relevant virtual content (what) at the right place (where) and the right time (when) with the attractive representation (how) is a significant and challenging problem in the field of virtual content insertion. As mentioned above, most existing related work and online services in practice have proposed various solutions to handle the problems about what, where, and when, but how to represent the virtual content is rarely touched and discussed. In other words, they focus on camera calibration, selection of insertion points and relevant advertisements, rather than representation of virtual contents.

Motivated by the above observations, we proposed and realized a novel system which constructs a visually relevant connection between the inserted virtual content and the source videos. The system vividly evolves the virtual contents along an incremental and interactive evolution process for inserting them into videos and provides users a creative way to enrich and enhance source video contents. Based on the concept of evolutionary biology, the videos are considered as not only carriers of message conveyed by the virtual content but also the environment in which the lifelike virtual contents live. The same as other virtual content insertion systems, the virtual contents could be the advertisements, trademarks, commercial logos, or any other images that users indented to insert. But the evolution of the virtual content would be divided into distinct yet dependent phases in which different evolutionary appearances and/or behaviors according to various interactions on the basis of predefined evolutionary mechanisms are generated in the proposed system. The evolution would be triggered to develop the virtual content from an organic cell to an intelligent being with the incremental stimuli induced from interactions. As a result, with the different source videos or the insertion points, the interactions between the virtual content and the source video would be completely different and produce an entertaining storyline accordingly.

The benefits of the induced evolutionary pathway in our system are twofold. First, the partial features of the inserted virtual content revealed in each phase would arouse audience’s curiosity and make them wonder about what the virtual content finally is. Second, it provides users various points of view to observe the virtual content and imperceptibly make it stick to viewers’ minds with enhanced impression. Through the interactive evolution, the virtual content changes its appearances and behaviors dramatically and improves the audience’s viewing experience to the original video in a less-intrusive manner. In addition, by creating interactions between the virtual content and the source video to construct their connections would make the augmented videos visually appealing and increase its acceptability and attractiveness.

The rest of this paper is organized as follows. In Section 2, we describe the essential ideas and introduce the proposed system. The video content analysis and the virtual content analysis are addressed in Section 3 and Section 4, respectively. The virtual content insertion is presented in Section 5. Section 6 shows the experiments and the evaluation results. Finally, the concluding remarks and the application scenarios are given in Section 7.

2. SYSTEM OVERVIEW

In this section, we describe the essential ideas of our work and then give an overview of the proposed system.

2.1 Essential Ideas

Undoubtedly, plain styles of static presentation for the virtual content insertion would let the storyline tend to be dull and boring. Therefore, it could not make augmented videos more entertaining and engaging. In addition, the inserted virtual content would be easily ignored or bring induced intrusiveness. Although [9] indicated that it could reduce the induced intrusiveness to viewer’s viewing experience by offering closely links between the virtual content and source video contents, the representation of virtual contents has not been studied adequately. Based on the above observation and inspired by the evolutionary biology, we create a novel representation which could be considered as a contextual relevance in terms of visual perception. Specifically, we assign characteristics to the virtual content and animate it by interacting with the video contents in a vivid way. Therefore, the videos are not only a message carrier but also a living environment for the lifelike virtual content, as shown in Figure 1.

We break up the evolution for the inserted virtual content into three distinct yet dependent phases, that is the cell phase, the microbe phase, and the creature phase, in which with different perception and behaviors. Arising from the external stimuli perceived in each phase, the appearance and behaviors of virtual content would change in an evolutionary way and reveal more and more colors and textures. Eventually, the virtual content becomes an intelligent creature and would interact with the source video contents, such as salient objects or attractive roles in videos.
In the beginning, the virtual content would drop into videos in a less-intrusive manner because of its simple appearance and behavior which are like a single-celled organism. Through the cell phase, the cell-like virtual content is subject to the background influence and acts passively. We assign the sense of touch to the virtual content as its initial sensor in the cell phase. With the sense of touch, the shape of the virtual content would be evolved according to the level of perceived stimuli. In other words, it would grow strong to accommodate itself to the background. Generally, the more disorder motion in the background, the more clutter in the region. Thereby, it would be less striking and feel safe if the shape of the virtual content changes quickly in the clutter region. Therefore, with more disorder fluid force induced from the background, the evolving would be faster. Note that the visual information about the virtual content that viewer could confirm is only its contour in the end of the cell phase.

Besides the sense of touch, the virtual content would have developed the sense of sight in the end of the cell phase. Therefore, the virtual content in the microbe phase would begin to discover and actively move close to the salient object with colorful appearance and develop the synonymous ability to enhance its own colors and textures. Each time the virtual content touches the colorful object it would absorb the color and texture and finally reveal the complete visual information.

In the creature phase, with the sense of bearing being matured, the virtual content would dance with the perceived aural stimuli, and interact with the moving salient object in an intelligent manner. In this way, it could construct a visually strong connection between virtual content and source videos and produce the entertaining storyline with uncertainty.

### 2.2 System Overviews

Figure 2 illustrates the overview of the proposed system, which consists of three stages: the virtual content analysis stage, the video content analysis stage, and the virtual content insertion stage. After selecting a video for inserting virtual content, both the visual and aural analyses are applied to the input video in the video content analysis stage. In the visual analysis, each frame of the input video is applied the frame profiling to estimate the motion information and discriminate the regions according to visual features. On the other hand, the ROI estimation module localizes the region of interest in each video frame by combining various visual features. In addition to the visual analysis, aural saliency analysis is performed to characterize the sound which accompanies the video. Finally, by combining the result of each module in the video content analysis stage, a multilayer feature space for each frame of the input video is automatically constructed. In the virtual content analysis stage, the characterization module is used to analyze the appearance of the virtual content for visually evolving in terms of shapes, colors and textures. With the information of characterization, the virtual content is assigned the various sensors and effectors for generating behaviors with evolutions in the behavior modeling module. Then, in the virtual content insertion stage, the animation generation module acts on the effectors and sensors of the inserted virtual content. Therefore, according to the features perceived by the sensors of the virtual content, the effectors automatically generate the corresponding reactions. In the layer composition module, the virtual layer, in which the virtual content is animated on, is integrated with the video layer. Eventually, the augmented video with virtual content overlay is produced. The detailed processes of each module in the proposed system are described in the following sections.

### 3. VIDEO CONTENT ANALYSIS

In the video content analysis stage, we analyze the input video and extract audiovisual features.

#### 3.1 Frame Profiling

For the purpose of simulating the input video as the living environment that the virtual content moves and behaves in, we define the motion activity as the source of background influence that could affect the movement of the virtual content. Then, we discriminate the regions in each video frame according to the colors and textures.

##### 3.1.1 Motion Estimation

We extract the motion vector, which is calculated based on the block-based motion estimation in the video coding, from the compression domain to achieve low-complexity. Therefore, we can efficiently estimate the motions in the videos. We denote the extracted motion vector at the \((x, y)\) coordinate of macroblocks in the \(n\)-th video frame by \(\text{MV}^n(x, y)\).

##### 3.1.2 Region Segmentation

In order to discriminate the influence induced from regions in the video frame, we use an unsupervised image segmentation algorithm, which is called JSEG [10], to segment each motion vectors embedded video frame into several disjoint color-texture regions. In this way, we construct the segmentation map for each video frame. The segmentation map of the \(n\)-th frame is denoted by \(S^n\) and defined as:

\[
S^n = \bigcup_{j} s^{n}_j, \quad \text{where } s^{n}_j \cap s^{n}_{j'} = \varnothing \text{ if } j \neq j',
\]

where \(s^{n}_j\) represents the \(j\)-th disjoint color-texture region in the \(n\)-th frame. Note that the number of segmentation regions may be different in each video frame. Figure 3 shows two examples of motion vector embedded video frame segmentations. In this work, each segmentation region is considered as an independent background with different influence to the virtual content and the region boundary would...
restrict or affect the movement of virtual contents according to the behavior modeling.

With the frame profiling module, we construct the background (BG) feature map, which describes the motion vectors and the segmented regions, for each video frame.

### 3.2 ROI Estimation

Based on different combinations of visual feature models, several approaches [11-12] are proposed to construct a saliency map by computing the attentive strength for each pixel or image block. On the basis of the saliency map, the region-of-interest (ROI) can be easily derived by evaluating the center of gravity and the ranging variance. In this work, four types of video-oriented visual features selected from low level features to high level features, i.e., contrast-based intensity, contrast-based color, motion, and human skin color, are adopted to construct corresponding feature maps independently. Therefore, using different weights to linearly combine the constructed feature maps can produce various saliency maps with different meanings. In our implementation, we construct two types of saliency map, i.e., HROI and LROI, which are defined by emphasizing the human skin color and contrast-based color, respectively, to distinguish the attractive salient region perceived by the virtual content in different phases. Figure 4 shows the ideas and the feature maps.

### 3.3 Aural Saliency Analysis

Based on the same idea proposed in [13], we define the aural saliency response \( AR(T_h, T) \) at a time unit \( T \) and within a duration \( T_h \), to quantify the salient strength of the sound. That is,

\[
AR(T_h, T) = \frac{E_{\text{av}}(T)}{E_{\text{av}}(T_h)} \cdot \frac{E_{\text{peak}}(T)}{E_{\text{peak}}(T_h)},
\]

where \( E_{\text{av}}(T) \) and \( E_{\text{peak}}(T) \) are the average sound energy and the sound energy peak in the period \( T \), respectively. In addition, \( E_{\text{av}}(T_h) \) and \( E_{\text{peak}}(T_h) \) are the maximum average sound energy and the maximum sound energy peak within duration \( T_h \), respectively. In order to suppress the noises from low frequencies, the sound energy is defined as the weighted sum over the spectrum power of an audio signal at a given time as follows:

\[
E = \int_0^{f_s} W(f) \cdot 10 \log(|SF(f)|^2 + \varsigma) df,
\]

where \( SF(f) \) is the short-time Fourier transform coefficient of the frequency component \( f \) and \( f_s \) denotes the sampling frequency. In addition, \( W(f) \) is the corresponding weighting function defined as follows:

\[
W(f) = \frac{1}{1 + e^{-f_1(f-f_2)}},
\]

where \( f_1 \) and \( f_2 \) are control parameters.

After analyzing the aural saliency of an audio, we construct an \( AR \) feature sequence which describes the aural saliency response at each time unit \( T \) in the video.

### 4. VIRTUAL CONTENT ANALYSIS

In the virtual content analysis stage, we analyze the visual features of the virtual content for developing the appearance evolving mechanism and define several motion styles to synthesize the lifelike behaviors.

#### 4.1 Virtual Content Characterization

In order to evolve the virtual content in terms of the appearance, we analyze its visual features, such as colors and textures. In this work, we use the prescribed color-texture segmentation algorithm JSEG to segment the virtual content image into several disjoint color-texture regions. Each region is defined as a cell of the virtual content. The segmentation map \( H \) for the virtual content is defined as follows:

\[
H = \bigcup_{i=1}^N h_i, \quad \text{where} \quad h_i \cap h_{i'} = \varnothing \quad \text{if} \quad i \neq i',
\]

where \( h_i \) is the \( i \)-th disjoint region in the segmentation map and \( N \) is the total numbers of regions. Figure 5 shows an example of the virtual content segmentation map.

In order to let the virtual content drop into the videos in a less-intrusively manner, the cells of the virtual content should be presented in sequence according to a suitable order. In this work, a region with the smallest area in the segmentation map is set as the initial cell to be presented and the next one should be the smallest one in its neighbors, which directly connect to the previous selected region, to avoid discontinuity. With the determined order, we can evolve the shape of the virtual content by controlling the

---

**Figure 3:** Two examples of motion vector embedded video frame segmentations. Left column shows video frame segmentations and right column shows the corresponding segmentation maps.

**Figure 4:** An example showing the LROI map (top-left) and the HROI map (top-right) in which high intensity parts indicate more attentive regions.
4.2.1 The Cell Phase

Initially, the BG sensor is activated to simulate the sense of touch and thereby the received surrounding force is the only one factor that affects the movement of the virtual content. With the sensor data, the TL effector updates the resultant force of the virtual content \( \mathbf{F}_{\text{vc}} \) by \( \mathbf{F}_s \). The regions for the virtual content moving around are restricted to its initial inserted segmentation regions.

The event for evolving is defined as that the change of direction of \( \mathbf{F}_{\text{vc}} \) is larger than a predefined degree. Once the event is detected by the CT sensor, the EV effector would control the opacity to present the cell of the virtual content; otherwise, the cell would be presented piece by piece in a predefined speed.

In order to reduce the visual intrusiveness, the EV effector harmonize the virtual content with its background (the video frame) based on the idea of [8] by using the method proposed in [14]. The harmonized colors of presented cells in each state are simplified to single color by averaging. Then, the contour of presented cell is darkened to enhance the shape information. Moreover, the techniques of morphing [15] is used to generate realistic transitions between two consecutive states, to further improve the effect of shape evolving, as shown in Figure 7. Once all the cells of the virtual content are presented, the virtual content would advance to the next phase.

4.2.2 The Microbe Phase

In this phase, the LROI sensor is additionally activated to simulate the sense of sight and the LROI region could be detected. In order to let the virtual content like to be attracted by the LROI region and moves forward to it, the TL effector updates the resultant force of the virtual content.
as follows:

\[
\begin{align*}
F_{ve} &= F_e(Q_H, R) + F_s \\
\angle(F_s) &= \tan^{-1}\left(\frac{|y_{Q_H} - y_R|}{x_{Q_H} - x_R}\right)
\end{align*}
\]  
where \((x_R, y_R)\) and \((x_{Q_L}, y_{Q_L})\) denote the centers of \(R\) and \(Q_L\) respectively, and \(Q_L\) represents the non-LROI region. Note that the resultant force of the virtual content \(F_{ve}\) would not be updated until an LROI region is detected. The virtual content is restricted to move in the \(Q_L\) region to avoid masking the salient region.

The event, in this phase, for triggering the evolving process is the collision between the virtual content and the LROI region. Each time the virtual content touches the boundary of an LROI region, the original colors of cell would be presented with the fading-in effect and the opacity would be modulated by a Gaussian function, as shown in Figure 8. At the same time, one of the segmentation regions in the LROI region would be visually erased by using the techniques of inpainting [16] to produce a background color overlay on it, as shown in Figure 9. In this way, the virtual content is simulated to obtain the color and texture by absorbing the energy of the salient object in the LROI region. Once the original colors of the virtual content are all presented, the virtual content would enter the final phase.

4.2.3 The Creature Phase

In the final phase, both the AR sensor and the HROI sensor are additionally activated to simulate the sense of hearing and develop a more penetrative sight. In other words, the virtual content is simulated to own the ability to dance with the perceived aural stimuli, and interacts with the moving salient object in an intelligent manner.

For the AR sensor data, we define two thresholds, say \(TH^A_L\) and \(TH^A_H\), for the RS effector to separately control the rotation and the scaling. If the AR sensor value is larger than \(TH^A_L\) but smaller than \(TH^A_H\), the RS effector computes a set of degrees of rotations to generate the jiggles effect. On the other hand, if the AR sensor data is larger than \(TH^A_H\), the RS effector computes the parameters of scaling to generate the effect of astonished expression.

For the HROI sensor data, the virtual content is simulated to begin to interact with moving salient regions and the TL effector updates the resultant force of the virtual content as follows:

\[
\begin{align*}
F_{ve} &= F_e(Q_H, R) + F_s \\
\angle(F_s) &= \tan^{-1}\left(\frac{|y_{Q_H} - y_R|}{x_{Q_H} - x_R}\right), \text{ if } F_e(Q_H) > TH^F \\
F_{ve} &= F_e(Q_H, R) + F_e(Q_H), \text{ otherwise}
\end{align*}
\]

where \(Q_H\) is the non-HROI region, \(TH^F\) is a threshold, and \((x_{Q_H}, y_{Q_H})\) is the center of HROI region \(Q_H\). The virtual content is restricted to move in the \(Q_H\) region to avoid masking the moving salient region. In this way, the salient region with different levels of motions would cause TL effector producing different reactions. Specifically, if the motion of salient object is smaller than the threshold \(TH^F\), the virtual content tends to imitate the behavior of the moving salient object. On the other hand, if the moving salient region behaves exaggeratedly, the virtual content would move to touch it.

In summary, the sensors and effectors described in this section can be considered as the genotype of the virtual content and different arrangements or parameter settings of sensors and effectors can generate different evolutions and interactions.

5. VIRTUAL CONTENT INSERTION

In the last stage, the system automatically generates impressive animations with evolutions on a virtual layer according to the video features and the virtual content behaviors. Finally, we overlay the virtual layer onto the input video to produce an augmented video with separated layers.

5.1 Animation Generation

After determining the inserted position, the animation generation module acts on the effectors and sensors of the inserted virtual content. The effectors begin to react to the features perceived by the sensors and automatically produce animations according the prescribed behavior settings in section 4. The received motion vectors are map to the force, as described in section 4, to quantify the influence induced from the videos. Specifically, the external force is computed as the weighted vector sum of motion vectors in the surrounding region of the \(n\)-th video frame, that is

\[
F(B^n, R^n) = \sum_{(x,y)\in B^n} G(x, y, x_{R^n}, y_{R^n}) \cdot \hat{M} \cdot V^n (x, y) \cdot I_{R^n}(x, y),
\]

where

\[
G(x, y, x_{R^n}, y_{R^n}) = e^{-\frac{(x-x_{R^n})^2+(y-y_{R^n})^2}{2\sigma^2}},
\]

\[
I_{R^n}(x, y) = \begin{cases} 
0, & (x, y) \in R^n \\
1, & \text{otherwise}
\end{cases}
\]

Note that \((x_{R^n}, y_{R^n})\) in eqn.(9) is the central macroblock in the region \(R\) of the \(n\)-th video frame, and \(I_{R^n}\) is used to
indicate whether the macroblock is in the region \( R \) of the \( n \)-th video frame. In addition, \( \bar{M}V^n(x, y) \) denotes that its direction is opposite to \( MV^n(x, y) \). Similarly, the internal force is computed as:

\[
F_0(Q^n) = \sum_{(x,y)\in Q^n} G(x, y, xQ^n, yQ^n) \cdot \bar{M}V^n(x, y). \tag{10}
\]

Accordingly, the displacement of the virtual content in the \((n + d)\)-th frame \( P^n(d) \) can be calculated by

\[
V^n(d) = V^n + \frac{F_{vc}^n}{M_{vc}} \cdot d, \tag{11}
\]

\[
P^n(d) = V^n \cdot d + \frac{F_{vc}^n}{2 \cdot M_{vc}} \cdot d^2, \tag{12}
\]

where \( V^n(d) \) is the simulated velocity of the virtual content in the \((n + d)\)-th frame, \( F_{vc}^n \) and \( M_{vc}^n \) represent the resultant force and the mass of virtual content in the \( n \)-th frame, respectively, and \( d \) is the duration of the \( n \)-th frame.

Note that the animations of virtual content are constructed on a virtual layer in the animation generation module.

\subsection{5.2 Layer Composition}

The final task of virtual content insertion stage is to produce the augmented video by integrating the virtual layer with the input video layer. The augmented videos with multiple layers are implemented by using the techniques of the Flash video. Therefore, the inserted virtual content could be hidden if users want to do. In addition, it can be easily to overlay various virtual contents with different behaviors on the same source video simultaneously for targeting the online audience based on demographic parameters.

\section{6. EVALUATION}

This section presents the evaluation of the proposed system. Since objective evaluation of the proposed system is difficult, subjective experiments were taken to investigate the performance of our system. There are eighteen subjects invited for the user study. We randomly select a video from Youtube as the source video and use a commercial logo which is contextually relevant to the selected video as the input to our system. We disabled all the sensors and only enable the TL effector to produce an augmented video called V1 with the animated virtual content inserted. On the other hand, we produced the augmented video, called V2, by setting all the sensors and effectors enable. The snapshots of the augmented video are shown in Figure 10.

The two augmented videos were presented to the subjects by projecting them onto a 60-inch screen and the subjects were asked to answer the following questions:

- Q1: Which one is relatively intrusive in 1\(^{st}\) phase?
- Q2: Which one is relatively impressive in each phase?
- Q3: Is V2 interesting in terms of the evolving shape in 1\(^{st}\) phase?
- Q4: Is V2 interesting in terms of the evolving colors and textures in 2\(^{nd}\) phase?
- Q5: Is V2 interesting in terms of the reaction to aural saliency in 3\(^{rd}\) phase?

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
Questions & Yes & No Diff. & No\
\hline
Q3 & 50.00\% & 50.00\% & 0.00\% \\
Q4 & 77.77\% & 16.67\% & 5.56\% \\
Q5 & 72.22\% & 16.67\% & 11.11\% \\
\hline
\end{tabular}
\caption{The results of questions about the entertaining interaction.}
\end{table}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{image10.png}
\caption{The snapshots of the augmented video. V2 is the augmented video with full features offered by our system.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{image11.png}
\caption{The evaluation results of Q2.}
\end{figure}

- Q6: Overall, can you accept the augmented video with interactive evolution produced by our system?

The evaluation result shows that only 16.7\% of the subjects think it is intrusive and 88.9\% of the subjects accept the augmented videos with interactive evolutions produced by our system. Figure 11 shows the percentage of agreements for the videos in Q2, it is apparent that the virtual content in V2 is very impressive in the second and the third phases, because of the rich and interesting behaviors and interactions. Although the inserted virtual content in the first phase is not impressive compared to other phases, most subjects commented that it arouse their curiosity and make them wonder about what the virtual content finally is. Furthermore, Table 1 shows that most subjects agree that the evolving effect and interactions in V2 are interesting.

In addition to the advertising videos, several movies were selected to be fed into our system for the purpose of entertainment. Several interesting interactions between the inserted virtual content and the role in the source video are automatically generated by our system, as shown in Figure 12. Note that the entertaining events which occur in
with the objects in videos for the purpose of virtual product placement.

8. REFERENCES


7. CONCLUSIONS

We have presented a novel system which provides a lifelike animation for inserting virtual content into general videos in a vivid way to enhance the impression and acceptability of virtual content. The evaluation results show that the produced videos improve the audience’s viewing experience to the original video content and engage viewers with interesting interactions. We believe that the proposed innovative system for the virtual content insertion would inspire users to create in-video entertaining storylines and turn their videos into visually appealing ones. In addition, it would bring a new opportunity to increase the advertising revenue for video assets of the media industry and online video services.

The application scenarios of our system are described in the following. First, humans may interact with the virtual content via the webcam, just like play an interactive game. Second, we provide a platform for the users to create and visualize the script of personal stories, such as the evolution process. Third, it’s a new tool for personal branding to build unique personal videos with originality. Finally, based on the concept of interaction, the virtual content would interact videos are not planned by users in advance. In other words, the occurred events only depend on the properties of videos and the predefined genotype of the inserted virtual content. Therefore, a new storyline, which is induced by a series of interaction events, can be generated automatically in the augmented video.