Visual Signatures in Video Visualization

Min Chen, Ralf P. Botchen, Rudy R. Hashim, Daniel Weiskopf, *Member, IEEE Computer Society*,
Thomas Ertl, *Member, IEEE Computer Society*, and Ian M. Thornton

Abstract— Video visualization is a computation process that extracts meaningful information from original video data sets and conveys the extracted information to users in appropriate visual representations. This paper presents a broad treatment of the subject, following a typical research pipeline involving concept formulation, system development, a path-finding user study, and a field trial with real application data. In particular, we have conducted a fundamental study on the visualization of motion events in videos. We have, for the first time, deployed flow visualization techniques in video visualization. We have compared the effectiveness of different abstract visual representations of videos. We have conducted a user study to examine whether users are able to learn to recognize visual signatures of motions, and to assist in the evaluation of different visualization techniques. We have applied our understanding and the developed techniques to a set of application video clips. Our study has demonstrated that video visualization is both technically feasible and cost-effective. It has provided the first set of evidence confirming that ordinary users can be accustomed to the visual features depicted in video visualizations, and can learn to recognize visual signatures of a variety of motion events.

Index Terms—Video visualization, volume visualization, flow visualization, human factors, user study, visual signatures, video processing, optical flow, GPU rendering.

1 Introduction

A video is a piece of ordered sequential data, and viewing videos is a time-consuming and resource-consuming process. *Video visualization* is a computation process that extracts meaningful information from original video data sets and conveys the extracted information to users in appropriate visual representations. The ultimate challenge of video visualization is to provide users with a means to obtain a sufficient amount of meaningful information from one or a few static visualizations of a video using $O(1)$ amount of time, instead of viewing the video using $O(n)$ amount of time where n is the length of the video. In other words, *can we see time without using time* (i.e., *showing and viewing images in sequence*)?

Video data is a type of 3D volume data. Similar to visualization of spatial 3D data sets, one can construct a visual representation by selectively extracting important information from a video volume and projecting it onto a 2D view plane. However, in many traditional applications (e.g., medical visualization), the users are normally familiar with the 3D objects (e.g., bones or organs) depicted in a visual representation. In contrast, human observers are not familiar with the 3D objects depicted in a visual representation of a video because one spatial dimension of these objects shows the temporal dimension of the video. The problem is further complicated by the fact that, in most videos, each 2D frame is the projective view of a 3D scene. Hence, a visual representation of a video on a computer display is, in effect, a 2D projective view of a 4D spatiotemporal domain.

Depicting temporal information in a spatial geometric form (e.g., a graph showing the weight change of a person over a period) is an abstract visual representation of a temporal function. We therefore call the projective view of a video volume an *abstract visual representation* of a video, which is also a temporal function. Considering that the effectiveness of abstract representations is well-accepted in many applications, it is more than instinctively plausible to explore the usefulness of video visualization, for which Daniel and Chen proposed the following three hypotheses [6]:

1. Video visualization is an (i) intuitive and (ii) cost-effective means of processing large volumes of video data.
2. Well constructed visualizations of a video are able to show information that numerical and statistical indicators (and their conventional diagrammatic illustrations) cannot.
3. Users can become accustomed to visual features depicted in video visualizations, or be trained to recognize specific features.

The main aim of this work is to evaluate these hypotheses, with a focus on visualizing motion events in videos. Our contributions include:

- We have, for the first time, considered video visualization as a flow visualization problem, in addition to volume visualization. We have developed a technical framework for constructing scalar and vector fields from a video, and for synthesizing abstract visual representations using both volume and flow visualization techniques.
- We have introduced the notion of *visual signature* for symbolizing abstract visual features that depict individual objects and motion events. We have focused our algorithmic development and user study on the effectiveness of conveying and recognizing visual signatures of motion events in videos.
- We have compared the effectiveness of four different abstract visual representations of motion events, including solid and boundary representations of extracted objects, difference volumes, and motion flows depicted using glyphs and streamlines.
- We have conducted a user study, resulting in the first set of evidence for supporting hypothesis (3). In addition, the study has provided an interesting collection of findings that can help us understand the process of visualizing motion events through their abstract visual representations.
- We have applied our understanding and the developed techniques to a set of real videos collected as benchmarking problems in a recent computer vision project [10]. This has provided further evidence to support hypotheses (1) and (2).

2 Related Work

Although video visualization was first introduced as a new technique and application of volume visualization [6], it in fact reaches out to

- *Min Chen and Rudy R. Hashim are with Computer Science, and Ian M. Thornton is with Psychology, Swansea University, UK; E-mails: {m.chen, csrudy, i.m.thornton}@swansea.ac.uk.*
- *Ralf P. Botchen and Thomas Ertl are with Visualization and Interactive Systems, University of Stuttgart, Germany; E-mails: {botchen, thomas.ertl}@vis.uni-stuttgart.de.*
- *Daniel Weiskopf is with GrUVi, Computing Science, Simon Fraser University, Canada; Email: weiskopf@cs.sfu.cs.*

Manuscript received 31 March 2006; accepted 1 August 2006; posted online 6 November 2006.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

1077-2626/06/$20.00 © 2006 IEEE Published by the IEEE Computer Society

a number of other disciplines. The work presented in this paper relates to *video processing*, *volume visualization*, *flow visualization*, and *human factors in motion perception*.

Automatic video processing is a research area residing between two closely related disciplines, image processing and computer vision. Many researchers studied video processing in the context of video surveillance (e.g., [4, 5]), and video segmentation (e.g., [18, 24]). While such research and development is no doubt hugely important to many applications, the existing techniques for automatic video processing are normally application-specific, and are generally difficult to adapt themselves to different situations without costly calibration.

The work presented in this paper takes a different approach from automatic video processing. As outlined in [25], it is intended to 'take advantage of the human eye's broad bandwidth pathway into the mind to allow users to see, explore, and understand large amounts of information at once', and to 'convert conflicting and dynamic data in ways that support visualization and analysis'.

A number of researchers have noticed the structural similarity between video data and volume data commonly seen in medical imaging and scientific computation, and have explored the avenue of applying volume rendering techniques to solid video volumes in the context of visual arts [9, 12, 15]. Daniel and Chen [6] approached the problem from the perspective of scientific visualization, and demonstrated that video visualization is potentially an intuitive and cost-effective means of processing large volumes of video data. Bennett and McMillan [1] also demonstrated that a spatiotemporal video volume can be used to aid the process of video editing.

Flow visualization is another important area in scientific visualization [16, 20, 26]. There exist several different strategies to display the vector field associated with a flow. One approach used in this paper relies on glyphs to show the direction of a vector field at a collection of sample positions. Typically, arrows are employed to encode direction visually, leading to hedgehog visualizations [7, 14]. Another approach is based on the characteristic lines, such as streamlines, obtained by particle tracing. A major problem of 3D flow visualization is the potential loss of visual information due to mutual occlusion. This problem can be addressed by improving the perception of streamline structures [13] or by appropriate seeding [11].

In humans, just as in machines, visual information is processed by capacity and resource limited systems. Limitations exist both in space (i.e., the number of items to which we can attend) [21] and in time (i.e., how quickly we can disengage from one item to process another) [19, 22]. Several recent lines of research have shown that in dealing with complex dynamic stimuli these limitations can be particularly problematic [3]. For example, the phenomena of change blindness [23] and inattentional blindness [17] both show that relatively large visual events can go completely unreported if attention is misdirected or overloaded. In any application where multiple sources of information must be monitored or arrays of complex displays interpreted, the additional load associated with motion or change (i.e. the need to integrate information over time) could greatly increase overall task difficulty. Visualization techniques that can reduce temporal load clearly have important human factors implications.

3 Concepts and Definitions

A *video* $\mathbf{V}$ is an ordered set of 2D image frames $\{I_1, I_2, \ldots, I_n\}$. It is a 3D spatiotemporal data set, usually resulting from a discrete sampling process such as filming and animation. The main perceptual difference between viewing a still image and a video is that we are able to observe objects in motion (and stationary objects) in a video. For the purpose of maintaining the generality of our formal definitions, we include motionlessness as a type of motion in the following discussions.

Let m be a *spatiotemporal entity*, which is an abstract structure of an *object in motion* and encompasses the changes of a variety of attributes of the object including its shape, intensity, color, texture, position in each image, and relationship with other objects. Hence the ideal abstraction of a video is to transform it to a collection of representations of such entities $\{m_1, m_2, \ldots, m_k\}$.

Video visualization is thereby a function, $F : \mathbf{V} \rightarrow I$, that maps a video $\mathbf{V}$ to an image I, where F is normally realized by a computational process, and the mapping involves the extraction of meaningful information from $\mathbf{V}$ and the creation of a visualization image I as an abstract visual representation of $\mathbf{V}$. The ultimate scientific aim of video visualization is to find functions that can create effective visualization images, from which users can recognize different spatiotemporal entities $\{m_1, m_2, \ldots, m_k\}$ 'at once'.

A *visual signature* $\mathscr{V}(m)$ is a group of *abstract visual features* related to a spatiotemporal entity m in a visualization image I, such that users can identify the object, the motion, or both by recognizing $\mathscr{V}(m)$ in I. In many ways, it is notionally similar to a handwritten signature or a signature tune in music. It may not necessarily be unique and it may appear in different forms and different context. Its recognition depends on the quality of the signature as well as the user's knowledge and experience.

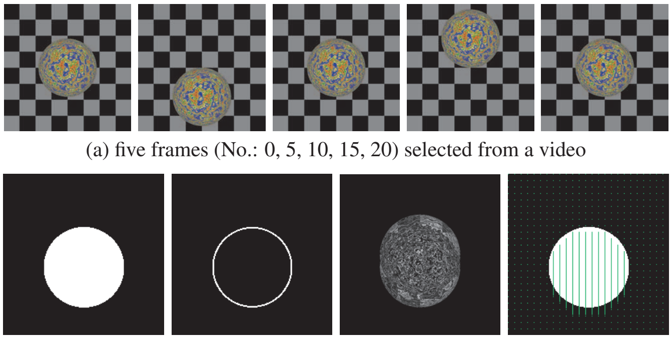
4 Types of Visual Signatures

Given a spatiotemporal entity m (i.e., an object in motion), we can construct different visual signatures to highlight different attributes of m. As mentioned in Section 3, m encompasses the changes of a variety of attributes of the object. In this work, we focus on the following time-varying attributes: (i) the shape of the object, (ii) the position of the object, (iii) the object appearance (e.g., intensity and texture), (iv) the velocity of the motion.

Consider an animation video of a simple object in a relatively simple motion. As shown in Fig. 1(a), the main spatiotemporal entity contained in the video is a textured sphere moving upwards and downwards in a periodic manner.

To obtain the time-varying attributes about the shape and position of the object concerned, we can extract the object silhouette in each frame from the background scene. We can also identify the boundary of the silhouette, which to a certain extent conveys the relationship between the object and its surroundings (in this simple case, only the background). Fig. 1(b) and (c) show the solid and boundary representations of a silhouette. To characterize the changes of the object appearance, we can compute the difference between two consecutive frames, and Fig. 1(d) gives an example difference image. We can also establish a 2D motion field to describe the movement of the object between each pair of consecutive frames, as shown in Fig. 1(e). There is a very large collection of algorithms for obtaining such attributes in the literature, and we will briefly describe our implementation in Section 6.

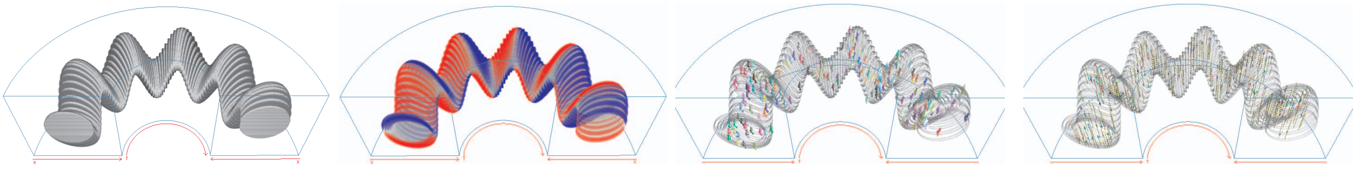
Compiling all silhouette images into a single volume results in a 3D scalar field that we call an *extracted object volume*. Similarly, we obtain an *object boundary volume* and a *difference volume*, which are also in the form of 3D scalar fields. The compilation of all 2D motion fields in a single volumetric structure gives us a *motion flow* in the form of a 3D vector field. Given these attribute fields of the spatiotemporal entity m, we can now consider the creation of different



(a) five frames (No.: 0, 5, 10, 15, 20) selected from a video

(b) silhouette (c) boundary (d) difference (e) motion flow

Fig. 1. Selected frames of a simple up-and-down motion, depicting the first of the five cycles of the motion, together with examples of its attributes associated with frame 1.



(a) Type A: silhouette hull (b) Type B: 4-band difference volume (c) Type C: motion flow with glyphs (d) Type D: motion flow with streamlines

Fig. 2. Four types of visual signatures of an up-and-down periodic motion given in Fig. 1.

visual signatures for *m*.

One can find numerous ways to visualize such scalar and vector fields individually or in a combinational manner. Without overcomplicating the user study to be discussed in Section 5, we selected four types of visualization for representing visual signatures. Each type of visual signature highlights certain attributes of the object in motion, and reflects the strength of a particular volume or flow visualization technique. All four types of visualization can be synthesized in real time, for which we will outline the technical framework in Section 6. For the following discussions, we chose the horseshoe view [6] as the primary view representation. In comparison with conventional viewing angles, it places four faces of a volume, including the starting and finishing frames, in a front view. It also facilitates relatively more cost-effective use of a rectangular display area, and conveys the temporal dimension differently from the two spatial dimensions.

4.1 Type A: Temporal Silhouette Extrusion

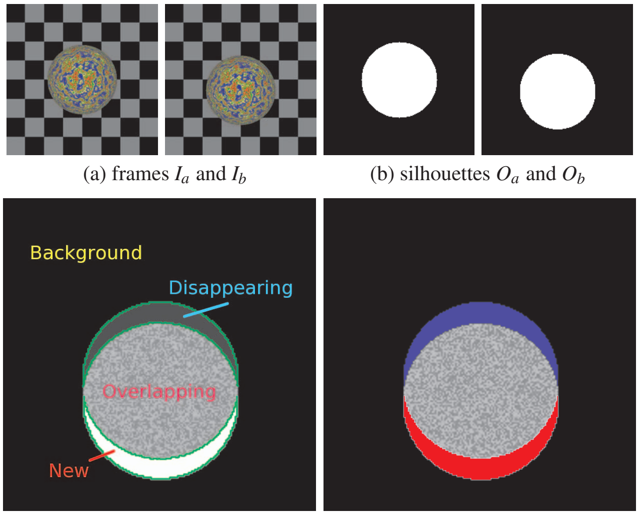
This type of visual signature displays a projective view of the *temporal silhouette hull* of the object in motion. Steady features, such as background, are filtered away. Fig. 2(a) shows a horseshoe view of the extracted object volume for the video mentioned in Fig. 1. The temporal silhouette hull, which is displayed as an opaque object, can be seen wiggling up and down in a periodic manner.

4.2 Type B: 4-Band Difference Volume

Difference volumes played an important role in [6], where amorphous visual features rendered using volume raycasting successfully depicted some motion events in their application examples. However, their use of transfer functions encoded very limited semantic meaning. For this work, we designed a special transfer function that highlights the motion and the temporal change of a silhouette, while using a relatively smaller amount of bandwidth to convey the change of object appearance (i.e., intensity and texture).

Consider two example frames and their corresponding silhouettes, O_a and O_b in Fig. 3(a) and (b). We classify pixels in the difference volume into four groups as shown in 3(c), namely (i) background ($\notin O_a \wedge \notin O_b$), (ii) new pixels ($\notin O_a \wedge \in O_b$), (iii) disappearing pixels ($\in O_a \wedge \notin O_b$), and (iv) overlapping pixels ($\in O_a \wedge \in O_b$). The actual difference value of each pixel, which typically results from a change detection filter, is mapped to one of the four bands according to the group that the pixel belongs to. This enables the design of a transfer function that encodes some semantics in relation to the motion and geometric change.

For example, Fig. 2(b) was rendered using the transfer function illustrated in Fig. 3(d), which highlights new pixels in nearly-opaque red and disappearing pixels in nearly-opaque blue, while displaying overlapping pixels in translucent gray and leaving background pixels totally transparent. Such a visual signature gives a clear impression that the object is in motion, and to a certain degree, provides some visual cues to velocity.



(a) frames I_a and I_b (b) silhouettes O_a and O_b

(c) 4 semantic bands (d) color mapping

Fig. 3. Two example frames and their corresponding silhouettes. Four semantic bands can be determined using O_a and O_b, and an appropriate transfer function can encode semantic meaning according to the bands.

4.3 Type C: Motion Flow with Glyphs

In many video-related applications, the recognition of motion is more important than that of an object. Hence it is beneficial to enhance the perception of motion by visualizing the motion flow field associated with a video. This type of visual signature combines the boundary representation of a temporal silhouette hull with arrow glyphs showing the direction of motion at individual volumetric positions. It is necessary to determine an appropriate density of arrows, as too many would clutter a visual signature, or too few would lead to substantial information loss. We thereby use a combination of parameters to control the density of arrows, which will be discussed in Section 6. Fig. 2(c) shows a Type C visual signature of a sphere in an up-and-down motion. In this particular visualization, colors of arrows are chosen randomly to enhance the depth cue of partially occluded arrows by improving their visual continuity.

Note that there is a major difference between the motion flow field of a video and typical 3D vector fields considered in flow visualization. In a motion flow field, each vector has two spatial components and one temporal component. The temporal component is normally set to a constant for all vectors. We experimented with a range of different constants for the temporal component, and found that a nonzero constant would confuse the visual perception of the two spatial components of the vector. We thereby chose to set the temporal components of all vectors to zero.

4.4 Type D: Motion Flow with Streamlines

The visibility of arrow glyphs requires them to be displayed in a certain minimum size, which often leads to the problem of occlusion. One alternative approach is to use streamlines to depict direction of motion flow. However, because all temporal components in the motion flow field are equal to zero, each streamline can only flow within the *x*-*y* plane where the corresponding seed resides, and it seldom flows far. Hence there is often a dense cluster of short streamlines, making it difficult to use color for direction indication.

To improve the sense of motion and the perception of direction, we mapped a zebra-like dichromatic texture to the line geometry, which moves along the line in the flow direction. Although this can no longer be considered strictly as a static visualization, it is not in any way trying to recreate an animation of the original video. The dynamics introduced is of a fixed number of steps, which are independent from the length of a video. The time requirement for viewing such a visualization remains to be $O(1)$. Fig. 2(d) shows a static view of such a visual signature. The perception of this type of visual signatures normally improves when the size and resolution of the visualization increases.

5 A User Study on Visual Signatures

The discussions in the previous sections naturally lead to many scientific questions concerning visual signatures. The followings are just a few examples:

- Can users distinguish different types of spatiotemporal entities (i.e., types of objects and types of motion individually and in combination) from their visual signatures?
- If the answer to the above is yes, how easy is it for an ordinary user to acquire such an ability?
- What kind of attributes are suitable to be featured or highlighted in visual signatures?
- What is the most effective design of a visual signature, and in what circumstances?
- What kind of visualization techniques can be used for synthesizing effective visual signatures?
- How would the variations of camera attributes, such as position and field of view, affect visual signatures?
- How would the recognition of visual signatures scale in proportion to the number of spatiotemporal entities present?

Almost all of these questions are related to the human factors in visualization and motion perception. There is no doubt that user studies must play a part in our search for answers to these questions. As an integral part of this work, we conducted a user study on visual signatures. Because this is the first user study on visual signatures of objects in motion, we decided to focus our study on the recognition of types of motion. We therefore set the main objectives of this user study as:

1. to evaluate the hypothesis that users can learn to recognize motions from their visual signatures.
2. to obtain a set of data that measures the difficulties and time requirements of a learning process.
3. to evaluate the effectiveness of the above-mentioned four types of visual signatures.

5.1 Types of Motion

As mentioned before, an abstract visual representation of a video is essentially a 2D projective view of our 4D spatiotemporal world. Visual signatures of spatiotemporal entities in real life videos can be influenced by numerous factors and appear in various forms. In order to meet the key objectives of the user study, it was necessary to reduce the number of parameters to be examined in this scientific process. We used simulated motions with the following constraints:

- All videos feature only one spherical object in motion. The use of a sphere minimizes the variations of visual signatures due to camera positions and perspective projection.
- In each motion, the center of the sphere remains in the same *x*-*y* plane, which minimizes the ambiguity caused by the change of object size due to perspective projection.
- Since the motion function is known, we computed most attribute fields analytically. This is similar to an assumption that the sphere is perfectly textured and lit, and without shadows, which minimizes the errors in extracting attribute fields using change detection and motion estimation algorithms.

We consider the following seven types of motion:

1. *Motion Case 1: No motion* — in which the sphere remains in the center of the image frame throughout the video.
2. *Motion Cases 2-9: Scaling* — in which the radius of the sphere increases by 100%, 75%, 50% and 25%, and decreases by 25%, 50%, 75% and 100% respectively.
3. *Motion Cases 10-25: Translation* — in which the sphere moves in a straight line in eight different directions (i.e., $0°, 45°, 90°, \ldots, 315°$) and two different speeds.
4. *Motion Cases 26-34: Spinning* — in which the sphere rotates about the *x*-axis, *y*-axis and *z*-axis, without moving its center, with 1, 5 and 9 revolutions respectively.
5. *Motion Cases 35, 38, 41: Periodic up-and-down translation* — in which the sphere moves upwards and downwards periodically in three different frequencies, namely 1, 5 and 9 cycles.
6. *Motion Cases 36, 39, 42: Periodic left-and-right translation* — in which the sphere moves towards left and right periodically in three different frequencies, namely 1, 5 and 9 cycles.
7. *Motion Cases 37, 40, 43: Periodic rotation* — in which the sphere rotates about the center of the image frame periodically in three different frequencies, namely 1, 5 and 9 cycles.

The first four types are considered to be elementary motions. The last three are composite motions which can be decomposed into a series of simple translation motions in smaller time windows. Five examples motion cases and their visual signatures can be found in the accompanying materials.

We did consider to include other composite motions, such as the periodic scaling, and combined scaling, translation and spinning, but decided to limit the total number of cases in order to obtain an adequate number of samples for each case while controlling the time spent by each observer in the study. We also made a conscious decision not to include complex motions such as deformation, shearing and fold-over in this user study.

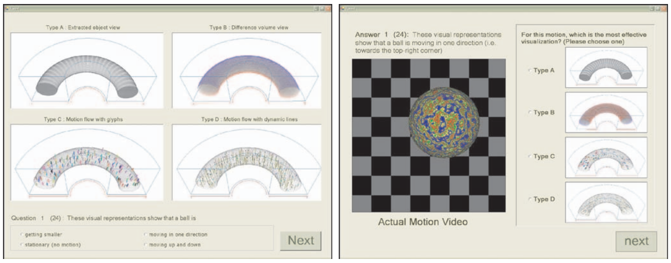
5.2 The Main User Study

Participants. 69 observers (23 female, 46 male) from the student community of Swansea University took part in this study. All observers had normal, or corrected to normal, vision and were given a £2 book voucher each as a small thank-you gesture for their participation. Data from two participants were excluded from analysis as their response times were more than 3 standard deviations outside of the mean. Thus, data from 67 (22 female, 45 male) observers were analyzed.

Tasks. The user study was conducted in 14 sessions over a three week period. Each session, which involved 4 or 5 observers, started with a 25 minutes oral presentation, given by one of the co-authors of this paper, with the aid of a set of pre-written slides. The presentation was followed by a test, typically taking about 20 minutes to complete. A piece of interactive software was specially written for structuring the test as well as collecting the results.

The presentation provided an overview of the scientific background and objectives of this user study, and gave a brief introduction to the four types of visual signatures, largely in the terminology of a layperson (see accompanying materials). It outlined the steps of the test, and highlighted some potential difficulties and misunderstandings. As part of a learning process, a total of 10 motions and 11 visual signatures were shown as examples in the slides.

The test was composed of 24 trials. On each trial, the observer was presented with between 1 and 4 visual signatures of a motion. As shown in Fig. 4(a), the task was to identify the underlying motion pattern by selecting from the 4 alternatives listed at the bottom of the screen. Both the speed and the accuracy of this response were measured. As observers were allowed to correct initial responses, the final



Question 1 (24) : These visual representations show that a ball is

○ getting smaller
○ stationary (no motion)
○ moving in one direction
○ moving up and down

Answer 1 (24): These visual representations show that a ball is moving in one direction (i.e. towards the top-right corner)

For this motion, which is the most effective visualization? (Please choose one)

○ Type A
○ Type B
○ Type C
○ Type D

Type A : Extracted object view
Type B : Difference volume view
Type C : Motion flow with glyphs
Type D : Motion flow with dynamic lines
Next
Actual Motion Video
next

(a) identifying motion patterns (b) feedback and evaluation

Fig. 4. Example screenshots of the main two tasks for each trial.

reaction time was taken from the point when they proceeded to the next part of the trial.

The second part of the trial was designed to provide feedback and training for the observers to increase the likelihood of learning. It also provided a measure of subjective utility, that is, how useful observers found each type of visual signature. In this part, the underlying motion clip was shown in full together with all four types of visual signatures (Fig. 4(b)). The task was to indicate which of the four visual signatures appeared to provide the most relevant information. No response time was measured in this part.

At the end of the experiment, observers were also asked to provide an overall usefulness rating for each type of visual signature. A rating scale from 1 (least) to 5 (most) effective was used.

Design. The 24 trials in each test were blocked into 4 equal learning phases (6 trials per phase) in which the amount of available information was varied. In the initial phase all 4 visual signatures were presented, providing the full range of information. In each successive phase, the number of available representations was reduced by one, so that in the final phase only one visual signature was provided. This fixed order was imposed so that observers would receive sufficient training before being presented with minimum information. For each observer a random sub-set of the 43 motion cases was selected and randomly assigned to the 24 experimental trials. For each case, the 4 possible options were fixed. The position of options was however randomized on the screen on an observer by observer basis to minimize simple response strategies.

5.3 The Supplementary User Study

Since the number of visual signatures available in the main user study decreased from one phase to another, it may be difficult to know whether changes in the overall accuracy and response times directly reflect learning. To address this issue, we conducted a supplementary user study, where two visual signatures, Types B and C, were made available throughout the 24 trials. It was organized in a same manner as the main study, and involved 40 observers (14 female, 26 male). Among them, 17 also took part in the main user study, hence had some experience of video visualization, with a time lapse of 4-5 months. The other 23 were first-time observers, with no previous experience in video visualization.

5.4 Results and Remarks

Analysis of Variance (ANOVA) was used to explore differences between three or more means, and t-tests were used to directly compare two means. By convention, F and t values indicate the ratio between effects of interest and random noise using specific probability distributions. The probability p of obtaining F or t values, given the statistical degrees of freedom indicated in parentheses, is also provided, with values less than 0.05 considered unlikely to occur by chance alone.

Motion Types. Table 1 gives the mean accuracy (in percentage) and response time (in second) in relation to motion types. There were clear differences between the types of motion, both in terms of accuracy ($F(4,264) = 34.5$, $MSE = 5$, $p < 0.001$), and speed ($F(4,264) = 12.6$, $MSE = 118$, $p < 0.001$).

The scaling condition gave rise to the highest accuracy, clearly showing that positive identification of motion is possible from visual signatures. Post-hoc analysis showed that this condition did not lead to better performance than the trivial static case, but performance was reliably higher than the other three motion types (all have $t > 6.0$, $p < 0.001$).

Accuracy levels for the translational motions, including the elementary motion in one direction, and combinational motion with periodical change of directions did not differ from each other, but were both significantly above those for spinning motion ($t > 2.8$, $p < 0.01$).

The difficulty in recognizing spinning motion appears to arise because the projection of the sphere in motion maintains the same outline and position throughout the motion. For example, the temporal silhouette hull of Motion Case 31, which is a spinning motion, is identical to that of Motion Case 1, which is motionless (see accompanying materials). This renders Type A visual signature totally ineffective in differentiating any spinning motion from the motionless state.

Table 1. Mean accuracy and response time related to motion types. Numbers in parentheses are standard errors (*se*) of the means.

	Accuracy (%)		Response time (second)	
Static	81.2	(4)	19.8	(2)
Scaling	90.3	(2)	13.6	(1)
Translation	66.7	(3)	23.8	(1)
Spinning	49.4	(3)	24.8	(1)
Periodic	62.2	(3)	24.4	(2)

Table 2. Mean accuracy and response time in each phase. The mean values are listed separately for the main user study, the first- and second-time groups in the supplementary user study. The standard errors (*se*) of the means listed are all between 1 and 2.

	Accuracy (%)			Response time (second)		
	main	sup-1	sup-2	main	sup-1	sup-2
Phase 1	66.7	*68.1*	*75.5*	30.8	*24.7*	*26.7*
Phase 2	70.0	*74.6*	*76.5*	22.2	*18.9*	*19.4*
Phase 3	72.0	*74.3*	*82.4*	17.5	*12.0*	*16.9*
Phase 4	63.0	*71.7*	*78.4*	13.4	*11.2*	*10.8*

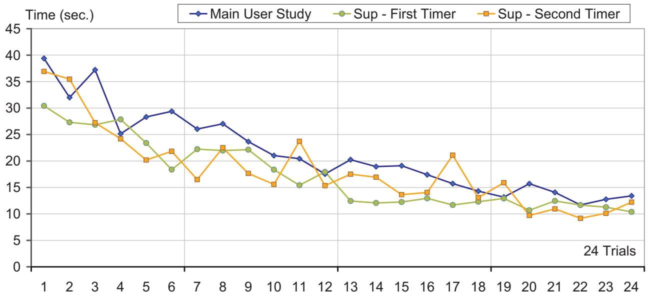


Fig. 5. The decreasing trend of the mean response time of each trial in both user studies.

Response times, computed only for correct trials, followed a similar pattern. Here, however, scaling motion did give rise to significantly better performance than the static case ($t(114) = 3.1$, $p < 0.001$), in addition to the other three moving cases. No other comparisons were significant.

Phases. Table 2 gives the mean accuracy (in percentage) and response time (in second) in each of the four phases. Although the supplementary study was not divided into specific phases, we grouped the data into 4×6 trials for comparison purposes.

In the main user study, accuracy levels changed significantly across the four phases ($F(3,198) = 2.9$, $MSE = 3.7$, $p < 0.05$). While there is a clearly increasing trend across the first 3 phases, this main effect appears to be due more to the final drop between phases 3 and 4, the only pair of means to differ significantly ($t(132) = 2.23$, $p < 0.05$). This drop may be due to the reduction of the number of visual signatures to only one in Phase 4. A single visual signature is often ambiguous, for example, spinning and static cases share the same Type A visual signature in our user studies, so this could have inflated error rates. Another possibility is the lack of a confirmation process based on a second visual signature.

We should note, however, that a similar trend can also be observed in the supplementary study, where Types B and C visual signatures were available throughout the session. Here, though, there was no main effect of phase. It seems possible that the generally high level of performance in both of the user studies may well be masking more subtle learning effects in terms of accuracy. Second time observers ($mean = 78\%$, $se = 2.6$) performed slightly better than first time observers ($mean = 72\%$, $se = 2.8$). Although this difference did not reach statistical significance, the trend towards higher performance is still

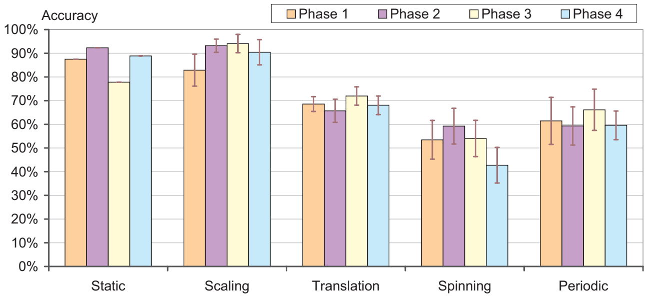


Fig. 6. The mean accuracy (with standard errors), measured in each of the four phases, categorized by the types of motion.

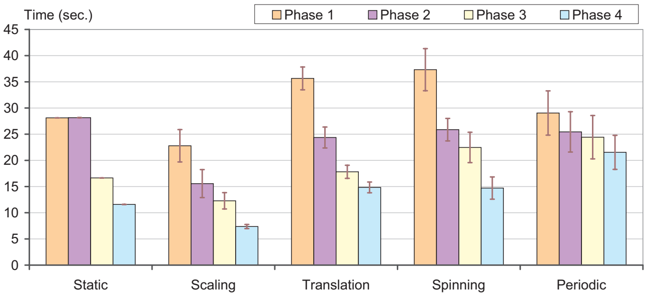


Fig. 7. The mean response time (with standard errors), measured in each of the four phases, categorized by the types of motion.

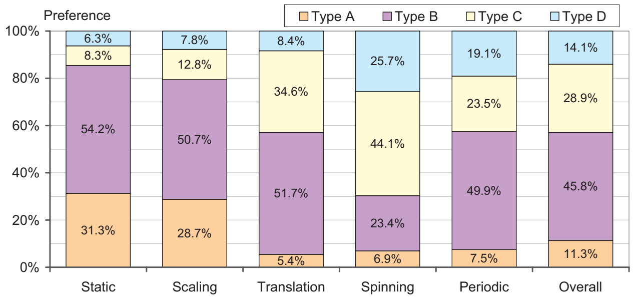


Fig. 8. The relative preference of each type of visual signature, presented in the percentage term, and categorized by the types of motion. The overall preference is also given.

encouraging. Any improvement, after a single prior exposure dating back several months, can provide some motivation to further explore long-term learning effects in this context.

In terms of response time, the story is much cleaner. In the main user study there was a clear effect of phase ($F(3,198) = 43.5$, $MSE = 97.8$, $p < 0.001$), which takes the form of a consistent linear decrease ($F(1,198) = 121.6$, $MSE = 97.8$, $p < 0.001$). Importantly, exactly the same pattern is present in the supplementary study, with a main effect of phase ($F(3,114) = 35.2$, $MSE = 45.1$, $p < 0.001$), driven by a linear decrease in response time ($F(1,114) = 103$, $MSE = 45.1$, $p < 0.001$). Thus, within the space of a single experiment, observers improve their performance even when the number of response options remains constant. There were no other significant response time effects in the supplementary study. Figure 5 shows this descreasing trend over the 24 trials for both user studies.

For the main study, Fig. 6 shows the accuracy in relation to each type of motion in each phase. We can observe that the spinning motion seems to benefit more from having multiple visual signatures available at the same time. The noticeable decrease of the number of positive identification of the motionless event in Phase 3 may also be caused by the difficulties in differentiating it from spinning. Fig. 7 shows a consistent reduction of response time for all types of motion.

Preference. Fig. 8 summarizes the preference of observers in terms of types of visual signatures, which largely reflects the effectiveness of each type of visual signature. Note that the Type C visual signature was considered to be the most effective in relation to the spinning motion, while Type B was generally preferred for other types of motion.

The overall preference (shown on the right of Fig. 8) was calculated by putting all 'votes' together regardless the type of motion involved. This corresponds reasonably well with the final scores, ranging between 1 (least) to 5 (most) effective, given by the observers at the end. The mean scores for the four types of visual signatures are A:2.6, B:4.0, C:3.6, and D:3.1 ($0.14 \leq se \leq 0.16$) respectively.

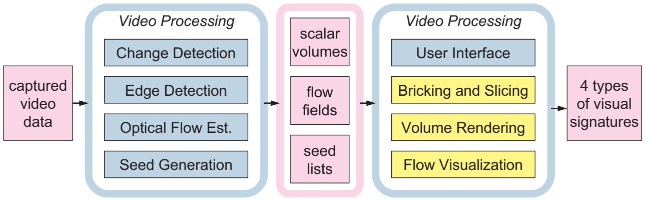


Fig. 9. The technical pipeline for processing video and synthesizing abstract visual representations. Data files are shown in pink, software modules in blue, and hardware-assisted modules in yellow.

6 SYNTHESIZING VISUAL SIGNATURES

Fig. 9 shows the overall technical pipeline implemented in this work. The main development goals for this pipeline were: (i) to extract a variety of intermediate data sets that represent attribute fields of a video. Such data sets include extracted object volume, difference volume, boundary volume, and optical flow field; (ii) to synthesize different visual representations using volume and flow visualization techniques individually as well as in a combined manner; and (iii) to enable real-time visualization of deformed video volumes (i.e., the horseshoe view), and to facilitate interactive specification of viewing parameters and transfer functions.

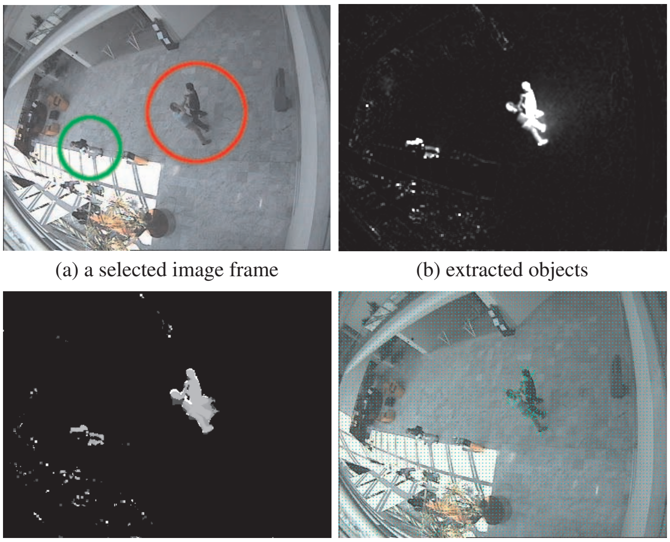
The *video processing* stage of the pipeline focuses on the generation of appropriate attribute fields, including *extracted object volume*, *4-band difference volume*, *object boundary volume*, *optical flow field*, and *seed list*. The *rendering* stage was implemented in C++, using Direct3D as the graphics API and HLSL as the GPU programming language. Volume rendering is based on 3D texture slicing. The flow visualization part is added by rendering opaque geometry that represents arrows or streamlines. For an 800 × 600 visualization and a 600 frame video, the volume renderer achieves about 12.9 fps on a 3.4GHz Pentium 4 PC with an NVIDIA GeForce 7800 GTX graphics board. Further details can be found in [2].

7 APPLICATION CASE STUDIES

We have applied our understanding and the developed techniques to a set of video clips collected in the CAVIAR project [10] as benchmarking problems for computer vision. In particular, we considered a collection of 28 video clips of the entrance lobby of the INRIA Labs at Grenoble, France, which were filmed from a similar camera position using a wide angle lens. Fig. 10(a) shows a typical frame of the collection, with actors highlighted in red, non-acting visitors in yellow. All videos have the same resolution with 384×288 pixels per frame and 25 frames per second. As all videos are available in compressed MPEG2, there is a noticeable amount of noise, which presents a challenge to the synthesis of meaningful visual representations for these video clips as well as automatic object recognition in computer vision.

The video clips recorded a variety of scenarios of interest, including people walking alone and in group, meeting with others, fighting and passing out, and leaving a package in a public place. Because the camera was located at a relatively high position and almost all motions took place on the ground, the view of the scene exhibits some similarity to the simulated view used in our user study. It is therefore appropriate and beneficial to examine the visual signatures of different types of motion events featured in these videos.

In this work, we tested several change detection algorithms as studied in [6], and found that the linear difference detection algorithm [8]



(a) a selected image frame

(b) extracted objects

(c) 4-band difference

(d) a computed optical flow

Fig. 10. A selected scene from the video 'Fight_OneManDown' collected by the CAVIAR project [10], and its associated attributes computed in the video processing stage.

is most effective for extracting an object representation. As shown in Fig. 10(b), there is a significant amount of noise at the lower left part of the image, where the sharp contrast between external lighting and shadows is especially sensitive to the minor camera movements, in addition to the noise caused by the lossy compression used in capturing these video clips. In many video clips, there were also non-acting visitors browsing in that area, resulting in more complicated noise patterns. Using the techniques described in Section 6 and [2], we also computed a 4-band difference image between each pair of consecutive frames (Fig. 10(c)), and an optical flow field (Fig. 10(d)).

Fig. 11 shows three different situations involving people leaving things around in the scene. Firstly, we can recognize the visual signature of the stationary objects brought into the scene (e.g., a bag or a box) in Fig. 11(b)-(e). In Type B, the part of motionless track appears to be colorless, while in Type C, there is no arrow associated with the track, indicating the lack of motion. In conjunction with the relative position and thickness of this part of the track, it is possible to deduce that an object is motionless on the floor.

We can also observe the difference among the three videos from their visualizations. In (c), the owner appeared to have left the scene after leaving an object (i.e., a bag) behind. Someone (in fact the owner himself) later came back to pick up the object. In (d), an object (i.e., a bag) was left only for a short period, and the owner was never far from it. In (e), the object (i.e., a box) was left in the scene for a long period, and the owner also appeared to walk away from the object in an unusual pattern.

Fig. 12 shows the visualization of two other video clips in the CAVIAR collection [10]. In the 'Fight_OneManDown' video, two actors first walked towards each other, then fought. One actor knocked the other down, and left the scene. From the visualization, we can identify the movements of people, including the two actors and some other non-acting visitors. We can also recognize the visual signature for the motion when one of the actor was on the floor as part of the track is associated with with very few arrows. This hence indicates the lack of motion. In conjunction with the relative position of this part of the track, it is possible to deduce that a person is motionless on the floor. We can observe a similar visual signature in part of the track in Fig. 12(c).

Visual signatures of spatiotemporal entities in real life videos can be influenced by numerous factors and appear in various forms. Such diversity does not in any way undermine the feasibility of video visualization, and on the contrary, it rather strengthens the argument for involving the 'bandwidth' of the human eyes and intelligence in the loop. The above examples can be seen as further evidence showing the benefits of video visualization.

8 Conclusions

We have presented a broad study of visual signatures in video visualization. We have successfully introduced flow visualization to assist in depicting motion features in visual signatures. We found that the flow-based visual signatures were essential to the recognition of certain types of motion, such as spinning, though they appeared to demand more display bandwidth and more effort from observers. In particular, in our field trial, combined volume and flow visualization was shown to be the most effective means for conveying the underlying motion actions in real-life videos.

We have conducted a user study that provided us with an extensive set of useful data about human factors in video visualization. In particular, we have obtained the first set of evidence showing that human observers can learn to recognize types of motion from their visual signatures. Considering that most observers had little knowledge about visualization technology in general, over 80% of them gained 50% or above success rate within a 45 minute learning process. The reduction of response time within a session is significant, while the improvement of accuracy may possibly gain through experiencing video visualization regularly over a period. Some of the findings obtained in this user study indicate the possibility that perspective projection in a video may not necessarily be a major barrier, since human observers can recognize size changes at ease. We are conducting further user studies in this area.

We have designed and implemented a pipeline for supporting the studies on video visualization. Through this work we have also obtained some first-hand evaluation as to the effectiveness of different video processing techniques and visualization techniques.

Acknowledgments

This work was partly supported by Royal Society UK for collaboration between Swansea and Stuttgart, and by an NSERC Discovery Grant to the fourth author. The treatment of videos as flows in visualization was inspired by Professor D. Silver (Rutgers, The State University of New Jersey), who in 2003 suggested studying optical flows in conjunction with video visualization.

References

[1] E. P. Bennett and L. McMillan. Proscenium: a framework for spatio-temporal video editing. In *Proc. ACM Multimedia*, pages 177–184, Berkeley, CA, 2003.

[2] R. P. Botchen, M. Chen, D. Weiskopf, and T. Ertl. GPU-assisted multi-field video volume visualization. In *Proc. International Workshop on Volume Graphics*, pages 47–54,135, 2006.

[3] P. Cavanagh, A. Labianca, and I. M. Thornton. Attention-based visual routines: Sprites. *Cognition*, 80:47–60, 2001.

[4] R. Chellappa. Special section on video surveillance (editorial preface). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):745–746, 2000.

[5] R. Cutler, C. Shekhar, B. Burns, R. Chellappa, R. Bolles, and L. Davis. Monitoring human and vehicle activities using airborne video. In *Proc. 28th Applied Imagery Pattern Recognition Workshop (AIPR)*, Washington, DC, 1999.

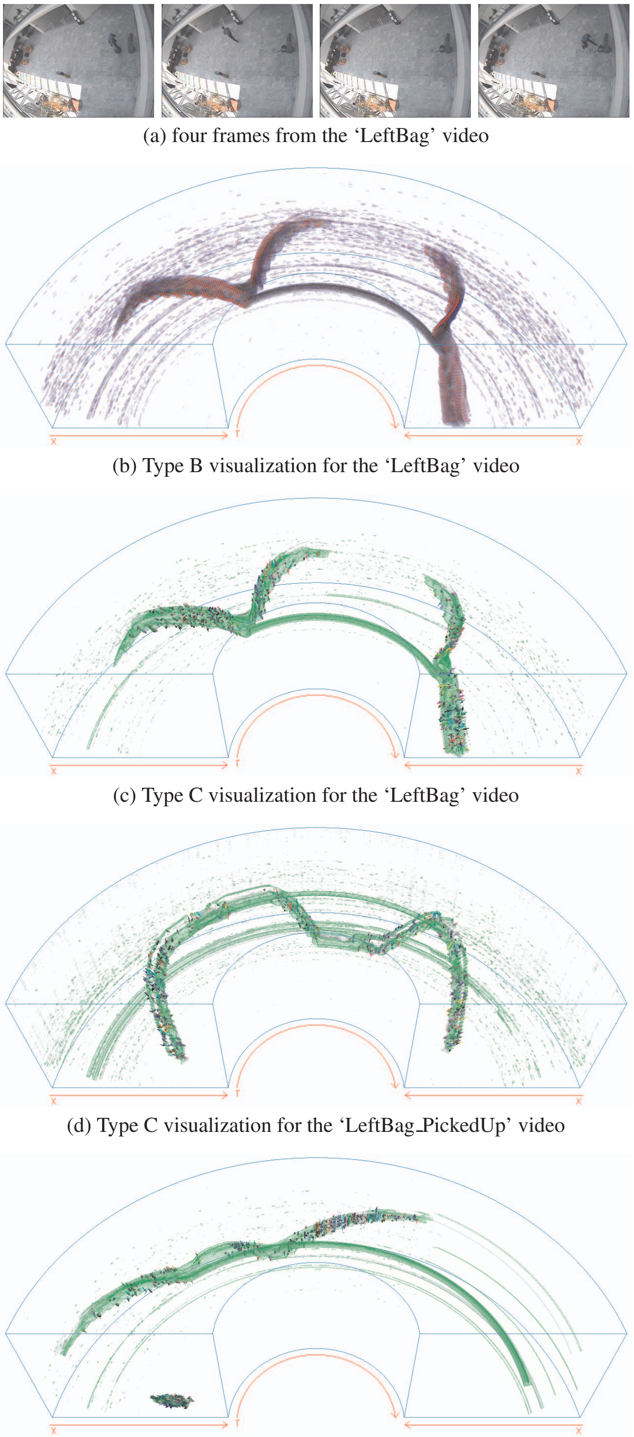
[6] G. W. Daniel and M. Chen. Video visualization. In *Proc. IEEE Visualization*, pages 409–416, 2003.

[7] D. Dovey. Vector plots for irregular grids. In *Proc. IEEE Visualization*, pages 248–253, 1995.

[8] T. E. E. Durucan. Improved linear dependence and vector model for illumination invariant change detection. In *Proc. SPIE*, volume 4303, San Jose, CA, 2001.

[9] S. Fels, E. Lee, and K. Mase. Techniques for interactive video cubism. In *Proc. 8th ACM International Conference on Multimedia (Posters)*, pages 368–370, 2000.

[10] R. B. Fisher. The PETS04 surveillance ground-truth data sets. In *Proc. 6th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 1–5, 2004.



(a) four frames from the ‘LeftBag’ video

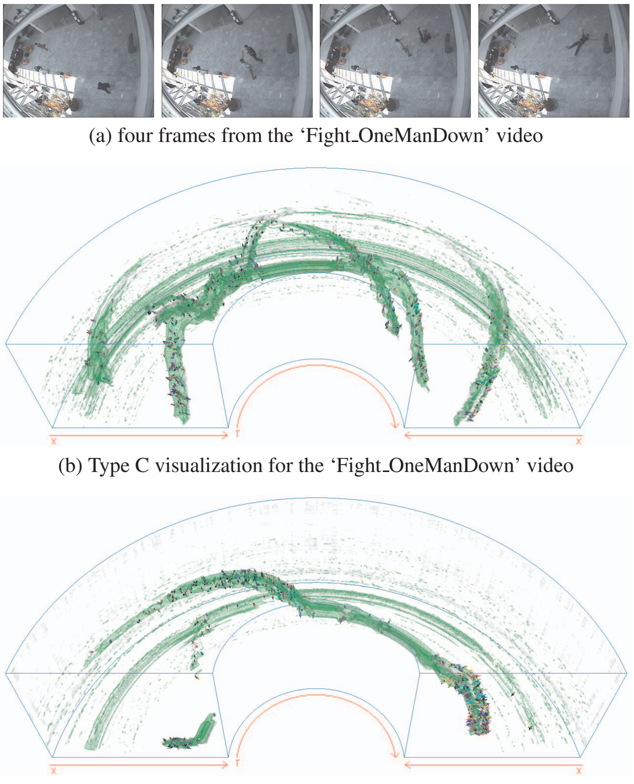
(b) Type B visualization for the ‘LeftBag’ video

(c) Type C visualization for the ‘LeftBag’ video

(d) Type C visualization for the ‘LeftBag_PickedUp’ video

(e) Type C visualization for the ‘LeftBox’ video

Fig. 11. The visualizations of three video clips in the CAVIAR collection [10], which feature three different situations involving people leaving things around. We purposely left out the original video frames for the ‘LeftBag_PickedUp’ and ‘LeftBox’ videos.



(a) four frames from the ‘Fight_OneManDown’ video

(b) Type C visualization for the ‘Fight_OneManDown’ video

(c) Type C visualization for the ‘Rest_SlumpOnFloor’ video

Fig. 12. The visualizations of two other video clips that feature situations involving people walking, stopping, and falling onto the floor.

[11] S. Guthe, S. Gumhold, and W. Straßer. Interactive visualization of volumetric vector fields using texture based particles. In *Proc. WSCG Conference Proceedings*, pages 33–41, 2002.

[12] A. Hertzmann and K. Perlin. Painterly rendering for video and interaction. In *Proc. 1st International Symposium on Non-Photorealistic Animation and Rendering*, pages 7–12, June 2000.

[13] V. Interrante and C. Grosch. Visualizing 3D flow. *IEEE Computer Graphics and Applications*, 18(4):49–53, 1998.

[14] R. V. Klassen and S. J. Harrington. Shadowed hedgehogs: A technique for visualizing 2D slices of 3D vector fields. In *Proc. IEEE Visualization*, pages 148–153, 1991.

[15] A. W. Klein, P. J. Sloan, R. A. Colburn, A. Finkelstein, and M. F. Cohen. Video cubism. Technical Report MSR-TR-2001-45, Microsoft Research, October 2001.

[16] R. S. Laramee, H. Hauser, H. Doleisch, B. Vrolijk, F. H. Post, and D. Weiskopf. The state of the art in flow visualization: Dense and texture-based techniques. *Computer Graphics Forum*, 23(2):143–161, 2004.

[17] A. Mack and I. Rock. *Inattentional Blindness*. MIT Press, Cambridge MA, 1998.

[18] N. V. Patel and I. K. Sethi. Video shot detection and characterization for video databases. *Pattern Recognition, Special Issue on Multimedia*, 30(4):583–592, 1997.

[19] M. I. Posner, C. R. R. Snyder, and B. J. Davidson. Attention and the detection of signals. *Journal of Experimental Psychology: General*, 109:160–174, 1980.

[20] F. H. Post, B. Vrolijk, H. Hauser, R. S. Laramee, and H. Doleisch. The state of the art in flow visualization: Feature extraction and tracking. *Computer Graphics Forum*, 22(4):775–792, 2003.

[21] Z. W. Pylyshyn and R. W. Storm. Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, 3:179–197, 1988.

[22] J. E. Raymond, K. L. Shapiro, and et al. Temporary suppression of visual processing in an RSVP task: An attentional blink. *Journal of Experimental Psychology, HPP*, 18(3):849–860, 1992.

[23] D. J. Simons and R. A. Rensink. Change blindness: past, present, and future. *Trends in Cognitive Sciences*, 9(1):16–20, 2005.

[24] C. G. M. Snock and M. Worring. Multimodal video indexing: a review of the state-of-the-art. *Multimedia Tools and Applications*, 2003.

[25] J. J. Thomas and K. A. Cook, editors. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Press, 2005.

[26] D. Weiskopf and G. Erlebacher. Overview of flow visualization. In C. D. Hansen and C. R. Johnson, editors, *The Visualization Handbook*, pages 261–278. Elsevier, Amsterdam, 2005.