2013 International Conference on Cyberworlds

ActVis: Activity Visualization in Videos

Masahiro Toyoura
University of Yamanashi
Kofu, Yamanashi, Japan
Email: mtoyoura@yamanashi.ac.jp

Satoshi Nishiguchi
Osaka Institute of Technology
Hirakata, Osaka, Japan
Email: nishigu@is.oit.ac.jp

Xiaoyang Mao
University of Yamanashi
Kofu, Yamanashi, Japan
Email: mao@yamanashi.ac.jp

Masayuki Murakami
Kyoto University of Foregin Studies
Kyoto, Japan
Email: masayuki@murakami-lab.org

***Abstract*—We present ActVis, which is a computer-aided video surveillance system for detecting and visualizing the activation levels of multiple objects in a video. ActVis indicates "something is happening" in a video. A user arranges panels indicating the regions of focusing objects on the video screen. Temporal differential as an activation level in a panel is detected by the system, and a corresponding seek bar representing the level is generated. In general, high-level features, such as body posture or facial direction/expression, cannot be extracted when the target object is partially occluded in video, or it is not human. By employing the temporal differential as a low-level feature and the metaphor of a level meter, our system can notify a user "when something happens." The user can explore high-level features of the moment. Potential applications of ActVis include the analysis of student activation levels in classroom for professional development of faculty, and observations of wild animals for ecological investigation.**

***Keywords*-video visualization; temporal differential; activation level; professional development of faculty; camera surveyllance;**

I. Introduction

Cameras are installed everywhere. The amount of captured videos is above the limit that we can skim through. Since auto recognition by machines is not perfect, we still have to skim the videos manually in many cases. Although shortcut keys of forwarding and rewinding are helpful as well as double speed playing, manual video skimming is a kind of exhausting task.

We aim to monitor the activation levels of multiple objects in a video. Taking the example of analysing lecture room video for professional development of faculty, the behaviour of students in the classroom is the most important information for evaluating the content and style of a lecture. By detecting the moment when students get distracted or starting interactions, one can gain the insights about how to prevent students from losing attentions or how to get them involved more in the class.

Motivated with such kind of applications, we propose a new computer-aided video surveillance system, called *ActVis* (**Act**ivity **Vis**ualization). Figure 1 shows the working window of the proposed system. ActVis adopts the metaphor of panel and seek bar for visualizing "something is happening" in real-time, and also provides a seek bar for allowing the user to explore "what is happening" in a particular frame in details. Panels are rectangular regions and can be interactively specified by a user with simple mouse clicks. They can be enveloping regions of focusing objects or some important locations to be monitored. The panels enables the users to detect the activity of a specified target even if a large portion of the target is occluded, without using predefined gestures and other rules.

Note that the panels are fixed on the target objects in the video. The video should be captured with a settled camera, and the target objects should not change their rough positions in the video.

In each panel, the amount of temporal differential between adjacent frames is detected as the activation level and visualized with a power meter placed at the left bottom corner of the panel. Unique IDs are assigned to the panels and indicated above the panels. The activation levels of a group or a whole can be monitored by combining differential of multiple panels and generating an integrated seek bar.

ActVis has the following 4 features which have not been realized in the previous systems described in the next section.

1) Effective visualization of the activation levels of multiple objects with the metaphor of panel and power meter.
2) Effective detection of events by adopting temporal differential as a low-level feature.
3) Indicating of "when it happens" rather than "what happened" for efficient video skimming.
4) Monitoring of multi-levels of activity, such as individuals, groups and the whole.

978-1-4799-2245-1/13 $31.00 © 2013 IEEE
DOI 10.1109/CW.2013.59

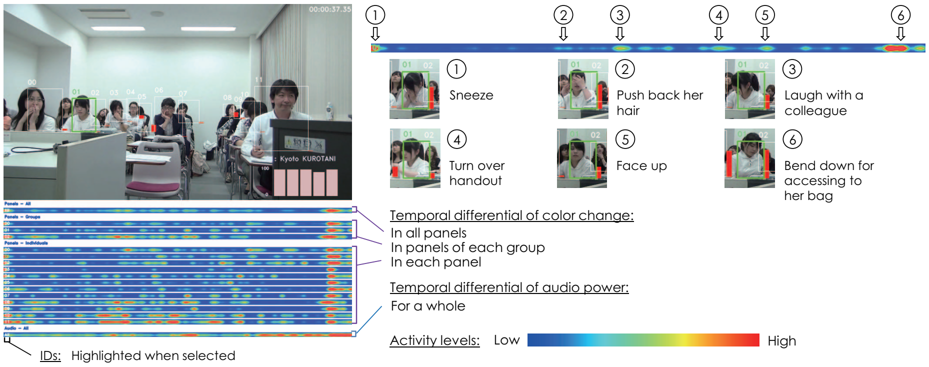
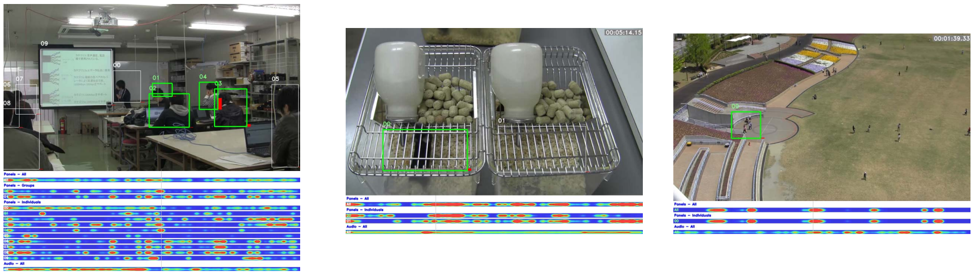


Figure 1. The activities of ID01 corresponding with the peeks in her seek bar.



(a) Lecture room from the back side.

(b) Mice in cages.

(c) Surveillance of a park.

Figure 2. Other applications of ActVis.

II. Related Works

A survey of video visualization has been published by Borgo et al. [3] It well covers recent works related to video visualization and video summaries. A pioneer work on video visualization was done by Daniel and Chen [6]. They visualized videos with horseshoe-shaped volumes. Important parts of the video are represented with color in the volume. A user can understand what happened in the video only with the volume data, although it requires the user to possess a special skill for checking it. Visual Signature [4] is another video visualization system which focuses on the transition of projective regions of moving objects. The differential area of the moving objects forms a colored tube in the horseshoe-shaped volume. Vis-a-Vis [9] visualizes the size of background differential in a video captured from a ceiling camera. The concept of Vis-a-Vis is similar to that of our ActVis. The activation level is measured by simple adjacent frame difference (AFD), and represented by a volume called activity cube. A major difference between ActVis and Vis-a-Vis is the level of detail. Vis-a-Vis aims to visualize the transition of activation level in the whole video, while ActVis aims to visualize the moment when the activation of a target raises and let a user to access arbitrary temporal position of the video. In addition, the concept of panel is first proposed in ActVis, which enable a user view the three levels of individuals, groups, and the whole.

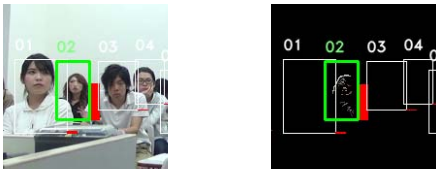
Video summary [8] is another solution to the visualization of videos. Video summaries are generated from selected frame images which are regarded to be important in well summarizing the whole story of the video. Video tapestries [2] gives a seamless summary of selected images. Dynamic video narrative [5] represents the trajectory of moving objects in an image. Although video summary enables the user to get a good overview of a video, it is not suitable for the applications where visualizing particular objects or groups of objects are required.

Automatic annotation for videos has also been proposed in the context of multimedia research. Annotations, or tags/indexes, enable users to overview a video. A video archive consists of mass of videos captured from massive cameras. Therefore annotations, such as who speaks, where a target moves, or which slide is presented, can facilitate efficient access to the archive and save time and effort of a viewer [7]. For videos on YouTube or other video-sharing sites, annotations by speech recognition and text recognition are provided for mining required videos [1]. These methods are based on pre-defined annotations of gestures, keywords, or other rules. The success rate of better annotating is dependent to how large or how clear the target appears without occlusion in the video. By using the temporal differential as the low-level feature, our technique can detect the activity of a specified target even if a large portion of the target is occluded, without using predefined gestures and other rules.

III. Activity Visualization

ActVis is based on the metaphor of “panel” and “power meter,” which is our newly proposed concept. Multiple targets in a video are visualized with their corresponding panels.

Figure 3 shows a single frame image of a sample video. From the original image without a panel and a power meter, we cannot recognize which target is in a high activation level. Then we generate an image shown in Figure 3(b). The bright pixels in the image represent there are large temporal differential at the pixels. We can recognize the target in panel ID02 moves at the moment. The red meter in the right of panel also visually indicates that there is a large movement in the panel.



(a) Original image (in zoom) (b) Temporal differential

Figure 3. Motion detection from temporal differential of panels.

In this case, the target in panel ID02 is small on the screen. Moreover the target is occluded by the other targets. Even if we employ a recent posture estimation method, the posture of the target would not be estimated correctly. By focusing temporal differential in the panel, we can detect the movement of the target.

Panel i is placed on the region R_i. Let us denote $I(x,y,t)$ as the color value at the position (x,y) of t-th frame. Whether there is large temporal differential is represented by the following function $f(x,y,t)$.

$$f(x,y,t) = \begin{cases} 0 & (||I(x,y,t)-I(x,y,t-1)|| < I_{th}) \\ 1 & (||I(x,y,t)-I(x,y,t-1)|| \geq I_{th}) \end{cases} \quad (1)$$

Note that I_{th} is a threshold for judging if there is large temporal differential. In experiment, we set I as the intensity and $I_{th} = 0.1$. Since the absolute difference value between colors of foreground and background does not have information, the difference should be binarized with I_{th}. Then, assuming $S(R_i)$ is the area of R_i, the activation level $v_i(t)$ for panel i is define by $f(x,y,t)$ as the follow.

$$v_i(t) = \sum_{(x,y)\in R_i} f(x,y,t)/S(R_i). \quad (2)$$

$v_i(t)$ is indicated by the power meter in the right of a panel i in real time. The meter visually represents the transition of $v_i(t)$. The length of the meter is controllable by users. The users can also choose the option of no meter.

A Seek bar for an individual panel indicates the transition of temporal differential in the corresponding panel. The temporal differential is visualized with a color map with blue standing for a low value and red for a high value. Seek bars for groups of panels and the all of panels are also generated for indicating the transition of activation levels of the groups and the whole.

In the leftmost of a seek bar, ID of the panel, group, or the whole is arranged. To visualize which seek bar is selected, selected ID and its corresponding panels are highlighted. For example in Figure 1, if a group consists of panels for ID 00 to 05, the panels are highlighted when the seek bar of the group is selected. When the seek bar of the whole is selected, all panels are highlighted.

If a user provides metadata corresponding to the panels, the metadata can be also displayed when the corresponding panel is selected. Figure 1 shows an example of it. As metadata, the names and results of past assessment of students are provided in this example. The function especially contributes to the online use of ActVis. A teacher can check metadata of the student in front of him/her.

For displaying seek bars under the video window, the width W of seek bars should be close to the width of the video image. The size of a video image is usually 10^2 to 10^4 pixels. Meanwhile, the number of video frames N is usually the order of 10^6 for a video of one hour. Then, we cannot simply plot $v_i(t)$ on a seek bar, otherwise the transition within a second cannot be represented on the seek bar. We set the maximum value in $t \in [Nx/W, N(x+1)/W)$ to sampling point $x \in [0, W-1]$ of the seek bar. Assuming temporal differential $v_i(t)$ are given at all frames, the pixel value $p_i(x)$ at the position x of a seek bar for panel i is calculated as follows:

$$p_i(x) \quad = \max_{t\in\left[\frac{N}{W}x,\frac{N}{W}(x+1)\right)} v_i(t). \quad (3)$$

Moreover, the values in the seek bar are normalized by the maximum value in the bar. The maximum value is given for each seek bar.

$$p_{i,max} \quad = \quad \max_{x} p_i(x). \quad (4)$$

Final pixel values are given as follows:

$$p_i(x) \quad \rightarrow \quad p_i(x)/p_{i,max}. \quad (5)$$

The seek bars for a group and the whole are generated in the same manner. For example, region R_{g_1} for a group g_1 is given as the union of regions $R_{g_{1-1}}, R_{g_{1-2}}, \cdots$ of the members $g_{1-1}, g_{1-2}, \cdots$ of g_1.

$$R_{g1} = R_1 \cup R_2 \cup \cdots. \quad (6)$$

The transition of loudness in the video is also visualized as a seek bar. A user can understand when something happened in the video, without reviewing the video. Current ActVis supports multiple channels, although we have sample videos with single channel loudness only. If there are multiple audio channels recorded from multiple microphones, multiple seek bars are generated for individual channels. When a group of microphones record the discussion of a group of students, the loudness of the microphones consists of a seek bar as well as panels. As a future work, we set up a well-calibrated system with multiple audio channels and confirm the validity the visualization with multiple seek bars.

IV. Experiments

Figure 1 shows an example of visualizing a lecture room video with ActVis. We focused on the panel for ID01 in Figure 1, and examined what happened where the corresponding seek bar indicated high values.

The action of facing up in Case 5 would be an important cue of students' attention for the lectures. We can estimate not only the status of the target student, but also the status of the class by reviewing multiple panels or groups, and the whole.

The action of push her hair back in Case 2 would be useless for the application of professional development of faculty. On the other hand, ActVis does not guarantee catching all important events for all applications. For example, although sleeping is an important event from the viewpoint of professional development, it may not be indicated in the seek bar and the meter on the panel. ActVis can visualize only those activities with high activation level. Nevertheless, ActVis would make possible to skim the activity of multiple targets in a long-term video.

Figure 2 shows other promising applications. Figure 2(a) shows an example of lecture video from the backside of a classroom. Students would feel more comfortable for being captured from the backside, but conventional estimators of human posture and facial expression cannot deal with such videos. In Figure 2(b), mice in cages were captured for analysing when they were in awaking state. In conventional systems, the state has been estimated with the sensors installed at the faucet of water bottles. ActVis contributed more detailed analysis. Figure 2(c) shows the video of an outdoor surveillance camera. A panel arranged on the entrance of the park indicates when and how many people come to and leave from the park. ActVis could detect the motion of such low-resolution humans. Temporal differential between adjacent frames is not affected by gradual illumination change.

V. Conclusions

We designed and implemented ActVis for visualizing the activation levels of multiple targets in videos. Panels on the screen enable us to detect “when something happened” in the video, while generated seek bars navigate a user to an arbitrary position of the video.

As a future work, we want to contribute to actual professional development of faculty with ActVis. For this purpose, more application tailored functions such a function for automatically creating a video clip by extracting the sequences of relevant events from the original video would be useful.

More detailed analysis for audio signals is another future work. We did not make videos with multiple audio channels, therefore we could not confirm the validity for videos with multiple channels. We will first construct a well-calibrated system for capturing a lecture and then validate the effectiveness of ActVis for videos with multiple audio channels.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 23300311.

References

[1] J. Adcock, M. Cooper, L. Denoue, H. Pirsiavash, and L. A. Rowe. Talkminer: a lecture webcast search engine. In *International Conference on Multimedia*, pages 241–250, 2010.

[2] C. Barnes, D. B. Goldman, E. Shechtman, and A. Finkelstein. Video tapestries with continuous temporal zoom. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, 25(3), 2010.

[3] R. Borgo, M. Chen, B. Daubney, E. Grundy, G. Heidemann, B. Höferlin, M. Höferlin, H. Leitte, D. Weiskopf, and X. Xie. State of the art report on video-based graphics and video visualization. *Computer Graphics Forum*, 31(8):2450–2477, 2012.

[4] M. Chen, R. Hashim, R. Botchen, D. Weiskopf, T. Ertl, and I. Thornton. Visual signatures in video visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1093–1100, 2006.

[5] C. D. Correa and K.-L. Ma. Dynamic video narratives. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)*, 29(3), 2010.

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

[7] A. Girgensohn, D. Kimber, J. Vaughan, T. Yang, F. Shipman, T. Turner, E. Rieffel, L. Wilcox, F. Chen, and T. Dunnigan. Dots: support for effective video surveillance. In *International Conference on Multimedia*, pages 423–432, 2007.

[8] A. G. Money and H. Agius. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation*, 19(2):121–143, 2008.

[9] M. Romero, J. W. Summet, J. T. Stasko, and G. D. Abowd. Viz-a-vis: Toward visualizing video through computer vision. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1261–1268, 2008.