

Low level Visio-Temporal Features for Violence Detection in Cartoon Videos

Tahira Khalil¹, Javed Iqbal²
Abasyn University Peshawar
Pakistan
tleo_22@yahoo.com
dr.javed.iqbal@abasyn.edu.pk

Awais Adnan
Institute of Management Sciences Peshawar
Pakistan
awaisadnan@gmail.com



ABSTRACT: Cartoons are an informative way for creating awareness; children take keen interest in watching cartoons and spend leisure time in front of television. Unfortunately there is an increasing trend of violence and other objectionable scenes in cartoon videos that have very bad impact on personality of young age groups. Extensive use of such violent scenes in media (movie and TV programs) is one of the factors of increase of violence in society. In this paper different low-level visual features are evaluated for violence detection in cartoon videos using indigenously developed dataset that is categorized as violent and non-violent. From our results it has been observed that low-level visual features could not be used efficiently for the detection of violence. However these features are very helpful in identifying the character and situation that if combined with a knowledgebase capability can be used for detection of violence and other objectionable elements in the cartoon clips.

Keywords: violence, cartoon, blood, explosion, color features

Received: 10 May 2016, Revised 17 June 2016, Accepted 27 June 2016

© 2016 DLINE. All Rights Reserved

1. Introduction

With advancement in communication technology the use of media has enhanced for entertainment purpose. Media has the potential to create impact on the viewer in both positive and negative way especially in case of children [1]. They mesmerize the child resulting in less social interaction and physical activities [2]. They can be educative, and informative but at the same time there are some factors in cartoons which have negative impact on child unconscious mind and give wrong image of what happens in our society [3]. In fact, many of these cartoons are harmful to kids due to violence, disobedience to elders, innuendo and bad language [4]. To protect children from violent content, manual annotation of cartoon videos is time consuming. Therefore means are require for automatic detection of violence content in cartoon videos.

Lot of work has been done for automatic violence detection in videos but most of these are developed for normal videos where

different low-level and high-level features are used in detection of violence detection. In case of cartoon videos, no considerable work could be found for automatic violence detection. With the exceptions of some scientific and mystery movies, most of the movies reflect real world environment. Cartoon movies on the other hand contain total virtual space developed in the imagination of its creator. Real videos mostly follow law of physics and have to maintain the original property for real world objects. In case of cartoon there is no such limitation, instead of the science it's the creativity of the developer that governs in cartoon movies. Because of these huge differences, most of the methods developed in real movies cannot be applied on cartoon films.

Cartoon videos are not natural videos, they are created using different means. They have simple texture, more colorful and have distinct black edges as compared to real time videos. In cartoons the hard realities of life are depicted in a lighter way due to which children in real life become less sensitive to any pain. Most of explosion, fighting and firing scenes are depicted in the form of fire with stars however in reality things are much different.

















Scene Types	Cartoon Videos		Real Videos	
Fighting				
Explosion				
Blood				
Guns				

Figure 1. Low-level Violence Features in Cartoon and Real Videos

So far many contents based detection systems have developed to identify and detect objectionable elements in videos. However most of them are based on high level features like shape and objects etc. Which are complex in nature and require computational time. Moreover this feature, because of their nature works better in real videos as compared to cartoon videos.

In this paper different low-level visual features are used to determine their importance in violence detection in cartoon videos. To conduct the experiment a dataset is first developed in which clips are divided into two categories, violent and normal cartoon film. Different low-level features are extracted and studied for both of the categories.

Paper is structured as follows. Section 2 provides literature review. Section 3, outline the Dataset & feature extraction. Results and discussions are presented in Section 4. Conclusion and future work is discussed in Section 5.

2. Literature Review

Cartoon is most favourite and popular program among children but with the passage of time content of cartoons have changed drastically. Some studies have shown that children who watch violence in cartoons are less sensitive to any pain.

They are so much used to it that in reality they take it as a lighter part and find fun in it. Lot of work has been done on content awareness and detection of violence in videos based on audio, video and combination of both features. S. Goto and T. Aoki [5] for violence detection in videos suggested mid-level violence clustering technique. It involve both audio-visual features and machine learning techniques. Multiple kernel learning is used to test both audio and visual modalities. System is trained and tested by dataset obtained from 2013 Affect Task and evaluated by MAP@100. Nam et al.[6] proposed multiple visual-audio features for violent scene detection, in which color tables were used for detection of blood and flames. Giannakopoulos et al. [7] presented framework based on visual-audio features for violence detection in movies and for decision of violent and non-violent videos k-Nearest Neighbor classifier is used. Derbas et al. [8] proposed Joint Audio-Visual Words representation, which constructs a codebook in the context of Bag-of-Words (BoW) by combining audio and visual features. Dai et al. [9] used Data from ImageNet and MIT scene dataset for detection of part-level attributes in each frame for identification of object. Combining them with other low-level features from both of visual and audio modalities, the SVM classifier is built. For detection of aggressive behavior Chen et al. [10] proposed a bag-of-words Framework and used binary local motion descriptors. Lin and Wang [11] for detection of violent scene used weakly-supervised audio violence classifier. Gong in [12] for detection of violent content used low-level audio-visual features and high-level audio effects. Clarin et al. [13] for detection of skin and blood pixel in each frame used Kohonen self-organizing map and for detecting violent actions motion intensity analysis is used. Vu Lam et al.[14] used multiple features for violent scene detection. Media Eval VSD 2014 dataset is used for subject research. Mid-level features are used for violent scene detection in videos. Similarly in [15] approach for detection of violent videos based on four visual features and for classification of violent and non-violent videos Hidden information SVM classifier is used. In [16] author used both local and global features for violence scene detection. For feature representation of each key frame Bag of words framework is used. Motion and audio features are also evaluated. Dataset for this research is obtained from from MediaEval Affect Task 2013. In [17] MediaEval 2015 violence dataset is enhanced by labelling videos manually based on ten subclasses. Using these subclasses and multimodal features SVM classifiers are trained to detect violence in videos.

In summary, all previous studies deal with low level, high level and multi modal features of violent detection in real time videos however according to existing knowledge no such work has been found for violence detection in cartoon videos.

3. Low-Level Features Extractions

To extract the low level features from cartoon videos two broad categories of visual information are used i.e Luma and Chroma. Luma are the brighter information which contain most of the data about video content. In case of cartoons these information play important role in understanding content of the scene. Most of the videos are in YPbPr format where Luma information is presented by Y channel. In this work videos are extracted in RGB format. From where the brightness information are extracted in the form of grey levels using following integer version of Gray-Scale conversion equation.

$$G_{(x,y)} = \frac{[R_{(x,y)} * 299 + G_{(x,y)} * 587 + B_{(x,y)} * 114] + 500}{1000} \quad (1)$$

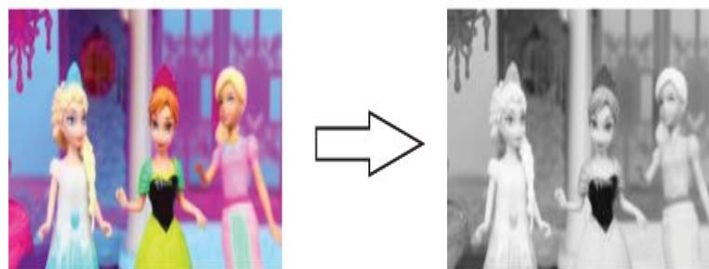


Figure 2. From RGB to gray-level

Where $G_{(x,y)}$ is the gray value at location x, y and $R_{(x,y)}$, $G_{(x,y)}$, $B_{(x,y)}$ are the red, green, blue values respectively of the pixel at corresponding location (x, y). This is the first integer version of the equation which gives near approximation. This process is summarized in the figure 2. below from this gray-level image 256—bin histogram is calculated from where 1st feature is

extracted as the peak of the vale in the histogram curve. In case of tie, first peak in the curve is selected as the feature. Initially mean value was selected as the feature but later experiments reviled that mean values do not represent the frame properly. As shown in the following figure 3 mean value is not true representative of the curve while the peak is a better option.

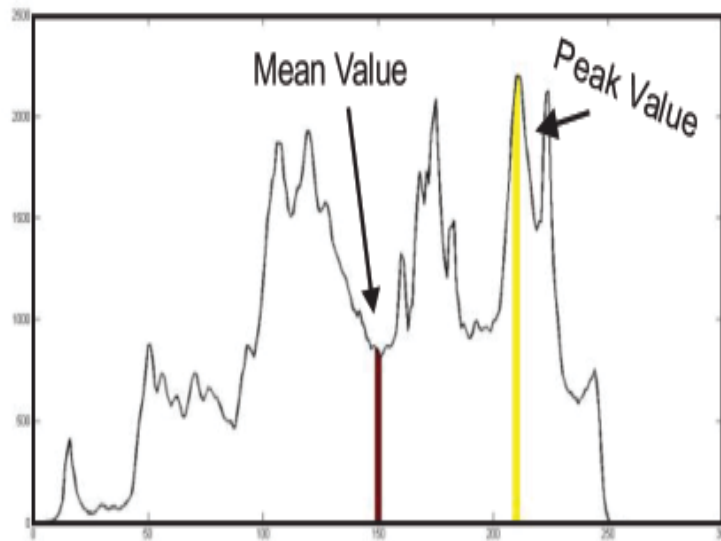


Figure 3. Mean vs peak value as feature selection

To extract information from Chroma domain two color models RGB and HSV are used. First frame extracted from the clip is converted into red, green and blue channels from where peak value from each channel is extracted. This conversion is summarized in the figure 4.

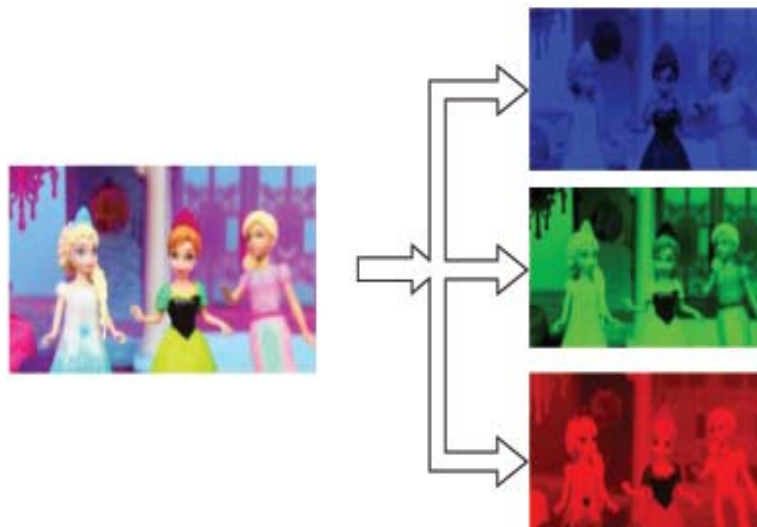


Figure 4. Conversion to Red, Green and Blue channels

Same frame is then converted into Hue, Saturation and value channel using following set of equations. Figure 5 summarized this step.

$$M = \max \{R, G, B\}$$

$$m = \min \{R, G, B\}$$

$$V = \frac{M}{255}$$

$$S = \begin{cases} 1 - m/M & \text{if } M > 0 \\ 0 & \text{if } M = 0 \end{cases}$$

$$H = \begin{cases} \cos^{-1} \left[\frac{R - \frac{1}{2G} - \frac{1}{2B}}{\sqrt{R^2 + G^2 + B^2 + -Rg - RB - GB}} \right] & \text{if } G \geq B \\ 360 - \cos^{-1} \left[\frac{R - \frac{1}{2G} - \frac{1}{2B}}{\sqrt{R^2 + G^2 + B^2 + -Rg - RB - GB}} \right] & \text{if } B > G \end{cases} \quad (2)$$



Figure 5. From RGB to HSV color space

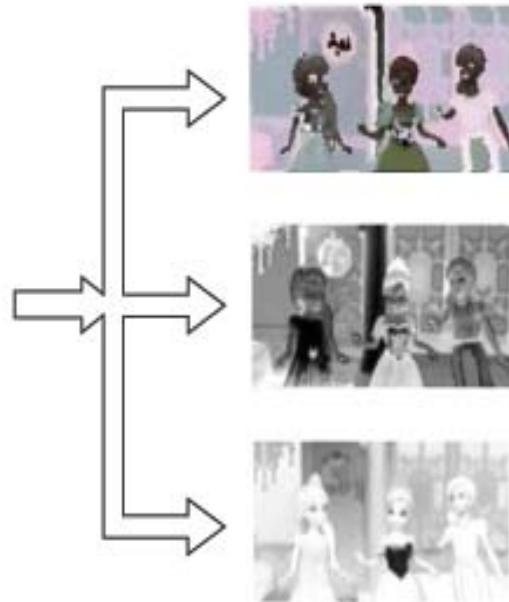


Figure 6. Example of a figure caption

Using these six channels peak for histogram is extracted for each frame and are thus combined as feature vector.

The whole process of this feature extraction is shown in figure 6.

Motion information is also used in this work to calculate the motion in the frame. For motion estimation frame is divided into 108 sub-blocks arranged into 9x12 arrays. Two dimensional motion vector is calculated using “Three-step Search (TSS) Algorithm for Block-Matching Motion Estimation Method”[22]. This method provides motion in Cartesian Coordinates which are then converted into Polar coordinates using the following equations.

Magnitude and the angle of the motion obtained from the above equation can be used to understand nature of the scene. Composite effect of these motion information is used as a feature in this work. Number of blocks with motion and their types are also used in feature vector.

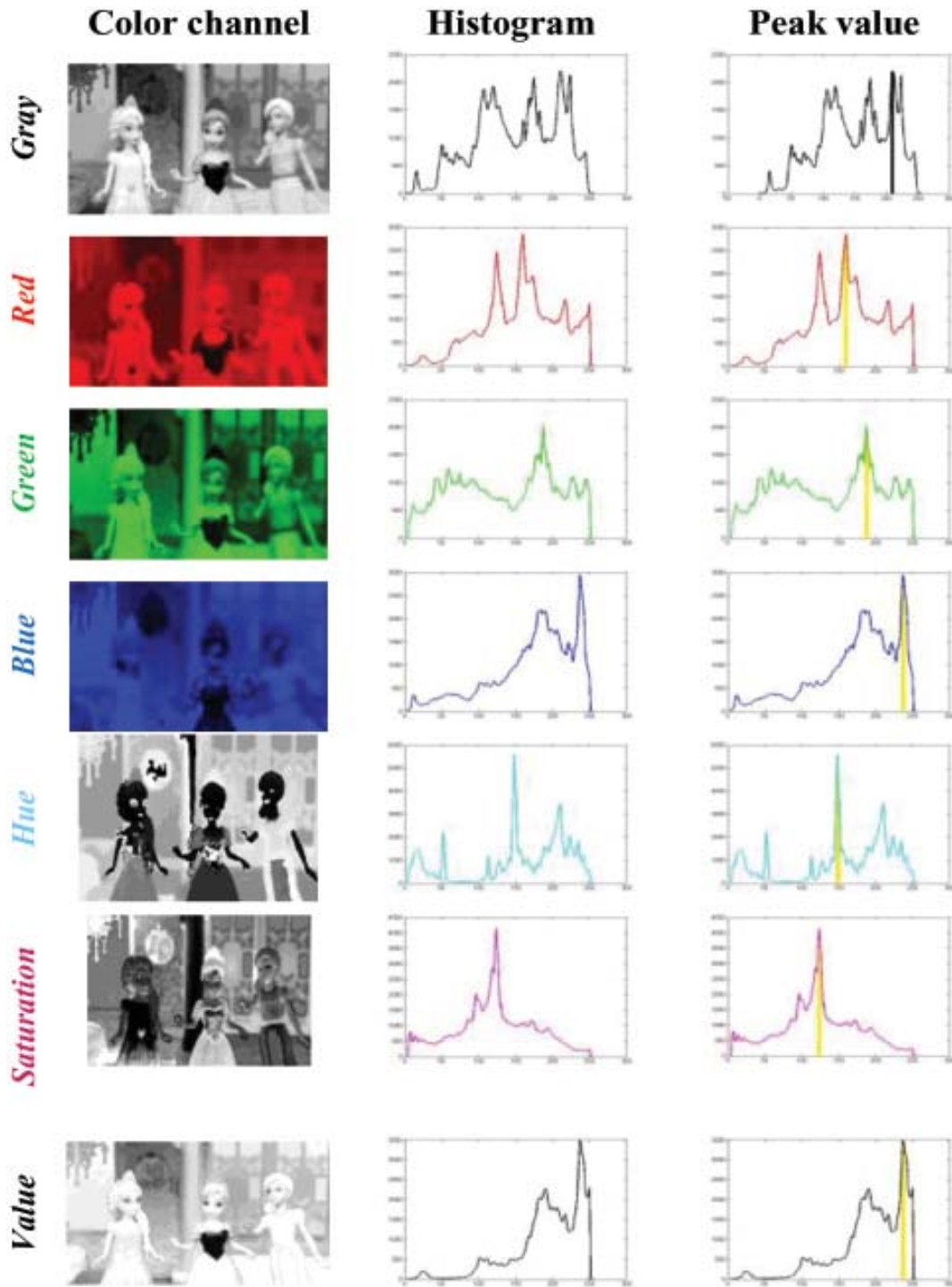


Figure 7(a). Feature extraction in seven channles

$$|g(j)| = \sqrt{(v_x(i,j))^2 + (v_y(i,j))^2} \quad (3)$$

$$\Theta_{(i,j)} = \tan^{-1}(v_y(i,j)/v_x(i,j)) \quad (4)$$



Figure 7(b). Feature extraction in seven channles

4. Dataset, Result And Discussions

4.1 Dataset

To analyze the relation between violence scenes and low level features, an indigenously created dataset has been used containing 16,654 seconds due to non-availability of cartoon dataset. Out of these 3,683 seconds of clips (11049 clips) contain violent content while rest have normal scenes. They contain total of 400175 frames. In preparation of dataset videos from different sources are collected which are then clipped into small segments of average length 60 mins and 1200 frames. These clips are then divided manually into violent and non-violent scenes. Summary of dataset is given in Table1.

S.No	Category	Number of clips	Number of frames	Duration (seconds)
1	Violent	112	11049	3683
2	Non-violent	392	389126	12971
	Total	504	400175	16654

Table 1. Summary of Dataset

4.2 Results

In most of content based image and video applications, high-level features are used, however in some cases low-level features also give considerable good results. RGB and HSV are two important color spaces for feature extraction.

In this work we have studied initially Red, blue feature from RGB color space and then Hues, Saturation and value from HSV color space. As illustrated in the figure xyz, no clear considerable difference could be found in the clips containing violence scenes and normal scenes. Next combination of these six channels with the brightness values was compared for the two categories. Results of this comparison are shown in figure xyz2. From the figure it can be seen that violence scenes have high value in Red channel. Similarly in Hue channel they lie in the band (). This is because of the fact that some of the violence scenes contain blood which forms patterns. However in cartoon videos most of the time such elements could be mixed with the other objects of red colors, especially in Barbie's videos.

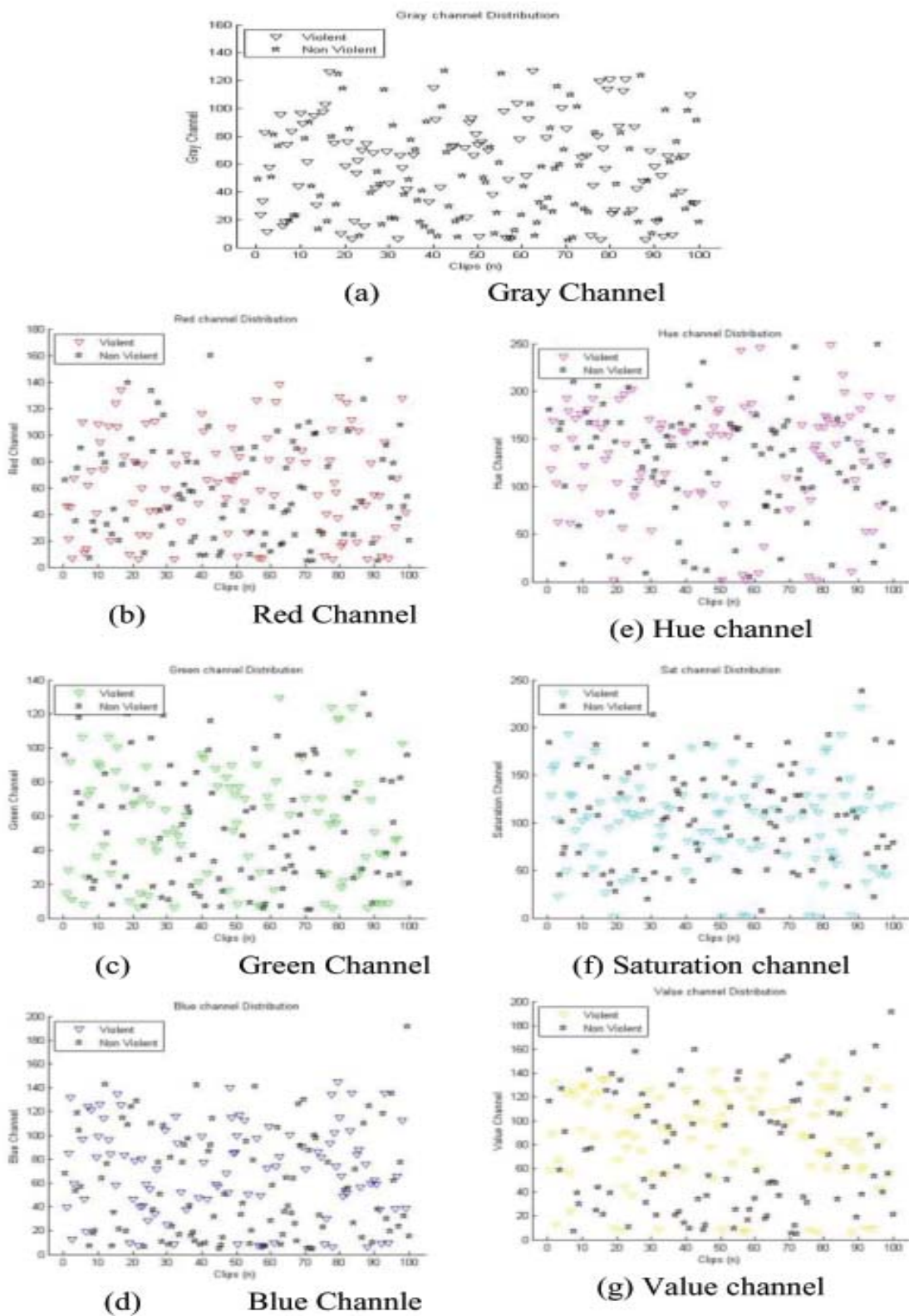


Figure 8. Features in RGB and HSV space for violent and non-violent visual clips in cartoon videos

References

- [1] Fouts, G., Callan, M., Piasentin, K., Lawson, A. (2006). Demonizing in Children's Television Cartoons and Disney Animated Films, *Child Psychiatry and Human Development*, 37 (1) 2006, p. 5-23.
- [2] Sudha, A. G. (2011). Factors Influencing The Change In Behaviour Of Children On Viewing Cartoon Programs- A Study, *Namex International Journal of Management Research*, 1 (1) 31-43.
- [3] Sultana, S. (2014). Role Of Cartoon: A Brief Discussion on How Cartoon put an Impact on Children, *ENH Community Journal* 1 (1) 9.
- [4] Kirsh, S. J. (2005). Cartoon violence and aggression in youth, *Aggression and Violent Behavior*, 11 (6) 547-557.
- [5] Goto S., Aoki, T. (2014). Violent Scenes Detection based on Automatically-generated Mid-level Violent Concepts, *In: 19th Computer Vision Winter Workshop Zuzana K'ukelov'a and Jan Heller (eds.) KČrtiny, Czech Republic, February 3-5, 2014.*
- [6] Nam, J., Alghoniemy, M., Tewfik, A. H. (1998). Audio-Visual Content-Based Violent Scene Characterization, *In: Proc. International Conference on Image Processing (ICIP) Vol. 1, Oct. 1998, p. 353-357.*
- [7] Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S., Theodoridis, S. (2010). Audio-visual fusion for detecting violent scenes in videos, *Artificial Intelligence: Theories, Models and Applications, 2010, p. 91-100.*
- [8] Derbas, N., Safadi, B., Quenot, G (2013). LIG at Mediaeval 2013 Affect Task: Use of a Generic Method and Joint Audio-Visual Words, *In: MediaEval Workshop, Oct. 2013.*
- [9] Dai, Q., Tu, J., Shi, Z., Jiang, Y. G., Xue, X. (2013). Fudan at MediaEval 2013: Violent Scenes Detection Using Motion Features and Part-Level Attributes. *In: MediaEval Workshop, Oct. 2013.*
- [10] Chen, D., Wactlar, H., Chen, M., Gao, C., Bharucha, A., Hauptmann, A (2008). Recognition of aggressive human behavior using binary local motion descriptors, *In: 30th Annual International Conference on IEEE Engineering in Medicine and Biology Society, Aug. 2008, p. 5238-5241.*
- [11] Lin, J., Wang, W. (2009). Weakly-supervised violence detection in movies with audio and video based cotraining, *In: Proceedings of the 10th Pacific Rim Conference on Multimedia, 2009, p. 930-935.*
- [12] Gong, Y., Wang, W., Jiang, S., Huang, Q., GAO, W (2008). Detecting violent scenes in movies by auditory and visual cues, *In: Proceedings of the 9th Pacific Rim Conference on Multimedia, 2008, p.317-326.*
- [13] Clarin, C., Dionisio, J., Echavez, M., Naval, P. C (2005). DOVE: Detection of movie violence using motion intensity analysis on skin and blood, *Technical Report, University of the Philippines, 2005.*
- [14] Lam, V., Phan, S., Le, D.D., Duong, D. A., Satoh, S (2016). Evaluation of multiple features for violent scenes detection, *Multimedia Tools and Applications, 2016, p. 1-25.*
- [15] Ji, X., Wu, O., Wang, C., Yang, J. (2014). Visual Feature-based violent video detection, *In: Proceedings of IEEE 3rd International Conference on Cloud Computing and Intelligence (CCIS), 2014, p. 619-623.*
- [16] Lam, V., DLe, D., Phan, S., Satoh, S., Duong, D. A., Ngo T. D. (2013). Evaluation of Low-Level features for detecting Violent Scenes in videos, *In: IEEE International Conference of Soft Computing and Pattern Recognition (SoCPaR), 2013, p. 213-218.*
- [17] Li, X., Huo, Y., Xu, J., Jin, Q. (2016). Detecting Violence in Video using Subclasses.
- [18] Lu., C., Liou, M. L (1997). A Simple and Efficient Search Algorithm for Block-Matching Motion Estimation, *IEEE Transactions on Circuits and Systems For Video Technology*, 7 (2) 429-433.