Semantic Video Analysis for Psychological Research on Violence in Computer Games

Markus Mühling[1], Ralph Ewerth[1], Thilo Stadelmann[1], Bernd Freisleben[1], Rene Weber[2], and Klaus Mathiak[3]

[1]Dept. of Math. & Computer Science
University of Marburg
Hans-Meerwein-Str.
D-35032 Marburg, Germany
{muehling, ewerth, stadelmann, freisleb}@informatik.uni-marburg.de

[2]Dept. of Communication
University of California Santa Barbara
Ellison Hall 4020
Santa Barbara, CA 93106, U.S.A.
renew@comm.ucsb. edu

[3] Dept. of Psychiatry & Psychotherapy
RWTH Aachen University
Pauwelsstr. 30
D-52074 Aachen, Germany
KMathiak@UKAachen.de

ABSTRACT

In this paper, we present an automatic semantic video analysis system to support interdisciplinary research efforts in the field of psychology and media science. The psychological research question studied is whether and how playing violent content in computer games may induce aggression. To investigate this question, the extraction of meaningful content from computer games is required to gain insights into the interrelationship of violent game events and the underlying neurophysiologic basis (brain activity) of a player. Previously, human annotators had to index game content according to the current game state, which is a very time-consuming task. The automatic annotation of a large number of computer game recordings (i.e. videos) speeds up the experimentation process and allows researchers to analyze more experimental data on an objective basis. The proposed computer game video content analysis system for computer games extracts several audiovisual low-level as well as mid-level features and deduces semantic content via a machine learning approach. This system requires manual annotations for a single video only to facilitate the semi-supervised learning process. Finally, human experts are allowed to refine the annotation results via a graphical user interface. Experimental results demonstrate the feasibility of the proposed approach.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Indexing methods.*

General Terms

Algorithms, Measurement, Experimentation, Human Factors.

Keywords

Semantic video analysis, computer games, psychology.

CIVR'07, July 9-11, 2007, Amsterdam, The Netherlands.

1. INTRODUCTION

Computer games play a very important role in today's entertainment media and belong to the most popular entertainment products. Unfortunately, the number of computer games containing serious violence increases. There is an extensive ongoing debate about the question whether playing violent games causes aggressive cognitions, aggressive affects or aggressive behavior, in particular with respect to teens and young adults.

The neurophysiologic perspective of mass communication research concentrates on emotional responses to video game playing. Mathiak and Weber [11] developed neurophysiologically grounded measures for the "human experience of media enjoyment". The study continues their prior work (Weber et al. [19]) on video game playing in which functional magnetic resonance imaging (fMRI) scans were taken during video game playing. Through this neurophysiologic perspective, they demonstrated that a specific neurological mechanism is activated when playing a first-person-shooter game. One central finding is that cognitive areas seem to suppress affective areas during the (virtually) violent interactions. This mechanism helps to better understand a potential link between playing certain types of violent video games and aggressive cognitions and affects.

The experimental design presented by Weber et al. [19] is based on the definition of certain game states and captures a player's brain activity via fMRI while (s)he is playing a violent computer game. Several semantic game events are distinguished: 1.) inactive; 2.) preparation; 3.) search and explore; 4.) danger; 5.) under attack, and 6.) fighting and killing. Once the game recordings are annotated with these semantic categories, the interrelationship of violent game events and the underlying neurophysiologic basis (brain activity) of the player can be investigated. Normally, human annotators are required to index such game content according to the current game state, but this is a very time-consuming task. In this context, computer-based automatic video content analysis of computer game recordings promises several advantages: Human annotation efforts can be reduced noticeably, and the annotation process is speeded up and is based on reproducible and objective criteria only. At the same time, researchers are enabled to investigate a larger number of computer game videos to gather more experimental data.

In this paper, we present an automatic semantic video analysis system that supports the experimental design described above by automatically identifying the game states (i.e. categories). The system is aimed at minimizing the human annotation effort and thus requires manual annotations for a single video only to facilitate the semi-supervised learning process. Content analysis relies on audiovisual low-level features as well as on mid-level features. The considered mid-level features are the results of shot boundary detection [3], camera motion estimation [4], audio segmentation, text detection [6] and face detection [13]. For each game category, a support vector machine (SVM) is trained using the low- and mid-level features. In our approach, only a single video sequence with a duration of 12 minutes is required to provide training data and hence, human annotation effort is kept at a minimum. Afterwards, new videos are automatically analyzed using these SVM models. To achieve a more robust result, an automatic semi-supervised correction step is employed separately for each video: Based on the initial classification result, the system automatically labels the frames in a new video and adapts its concept models to this video by employing feature selection and adaptively building a specialized classifier for a particular game video. Finally, the graphical user interface of our software system *Videana*[1] allows a human expert to refine respectively correct the annotation results, if needed. Experimental results demonstrate the very good performance of the proposed approach, which is thus indeed applicable to this interdisciplinary research field.

The paper is organized as follows. In section 2, related work of semantic analysis of videos for certain genres is discussed. Section 3 presents the experimental design and the main processing steps of the proposed semantic video analysis system. Experimental results are presented in section 4. Section 5 concludes the paper and outlines areas for future work.

2. RELATED WORK

To the best of our knowledge, neither video content analysis methods have been applied to computer game recordings nor automatic video content analysis has been suggested for the field of behavioral sciences. Nevertheless, there exist many semantic video analysis systems which are specialized for a certain genre, e.g. sports videos or news videos.

There are many approaches addressing the analysis of news videos. This emphasis might have been enforced by the TRECVID evaluation series [17] in which comprehensive news video test collections have been provided and used for evaluation purposes. A summary of semantic concept detection approaches regarding news videos is presented by Naphade and Smith [12]. The authors state that in most approaches, concept detection is considered as a supervised pattern recognition problem.

In a way, sports videos can be considered as somewhat related to the genre of computer games investigated in this paper: Since both genres are rule-driven, the amount of possibly appearing content is limited in both sports and computer games ("e-sports"). The automatic indexing of sports videos has been extensively studied in recent years. As noted in [15], many specific approaches exist for several sports domains, e.g. Formula-1, cricket, tennis, American football, and Gaelic football.

Apart from specific approaches, frameworks have been proposed that cover more than only a single type of sports. For example, Xu and Chua [21] propose a framework for event detection in team sports videos that is based on audiovisual features, domain knowledge, and external information sources.

Tong et al. [16] suggest a framework for semantic shot representation of sports videos. This framework is applicable to field sports, and shots are classified based on camera distance, displayed subject and edited video layout.

Sadlier and O'Connor [15] present an event detection system for field sports as well. They argue that it is not feasible to build a generic supervised event detection system for any kind of sports and find the limitation to field sports reasonable. The following features are employed in a supervised learning process: image crowd detection, speech-band audio activity, on-screen graphics tracking, motion activity measure, field line orientation and some other features.

3. SEMANTIC ANALYSIS OF COMPUTER GAME VIDEOS

In this section, we present our system to support interdisciplinary research in media and behavioral sciences via automatic multimodal video content analysis. First, in section 3.1 we describe the semantic classes which must be recognized for the experiment conducted by Weber et al. [19]. Then, our system is presented in sections 3.2 to 3.4. It utilizes automatically extracted audiovisual low-level and mid-level features to infer about the semantic game classes via supervised learning respectively semi-supervised learning. We have pursued two main targets. First, the system is supposed to remain a generic video content indexing system and thus does not contain any specific content detectors (restricting its applicability to a certain computer game would offer a lot of tuning possibilities). Second, the annotation effort that is needed to apply a machine learning approach should be kept at minimum, i.e. we allow the system to use a single labeled training video only. The following parts of our system are discussed in more detail in sections 3.2-3.4: audiovisual feature extraction, feature selection, classification, and a semi-supervised classification approach.

3.1 Semantic Classes for the Computer Game Experiment

Participants of the experiment conducted by Weber et al. [19] played the "mature" rated first-person-shooter game "Tactical Ops: Assault on Terror" [http://www.tactical-ops.de/]. As mentioned above, the experiment was aimed at gaining insight into the interrelationship of playing violent computer games and changes in the consumer's brain activities. Therefore, several game states were defined, and the dependence of the players' brain activity is set in relation to these game states. Brain activity was measured via fMRI scans. In this section, we present a system that is able to classify the following semantic classes with an acceptably high accuracy:

[1] Our work on video content analysis and retrieval is motivated by a large media research project currently conducted at the Universities of Siegen, Marburg, Dortmund, and FHG St. Augustin, Germany, entitled "Media Upheavals". The goal of our subproject is to provide a high-performance video content analysis system to support other subprojects applying film analysis. The software system Videana is currently under development to provide such support.

1.) “inactive”: The player’s avatar (PA) is dead or the game has not started yet.

2.) “preparation”: The PA is buying equipment in the beginning of a new round.

3.) “search/explore/danger”: The PA explores the virtual world and searches for hostages, enemies and weapons.

4.) “violence”: The PA is fighting and/or injured.

In the original study, the semantic game categories were distinguished and annotated more sophistically (see figure 1).

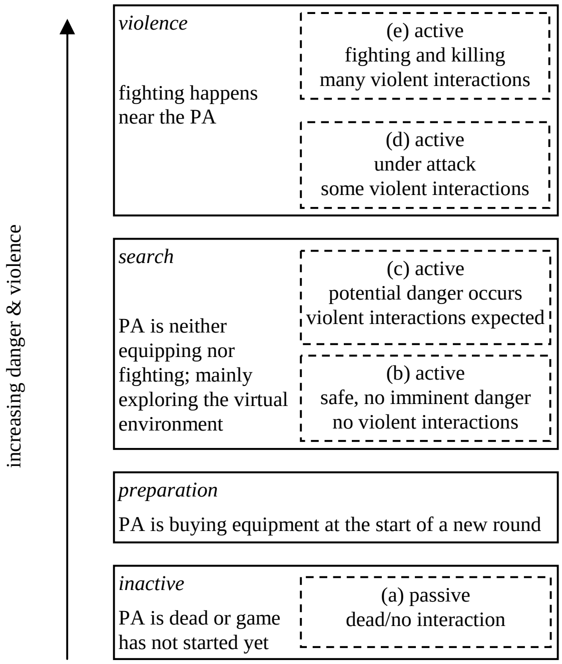
increasing danger & violence
violence
fighting happens near the PA
(e) active
fighting and killing
many violent interactions
(d) active
under attack
some violent interactions
search
PA is neither equipping nor fighting; mainly exploring the virtual environment
(c) active
potential danger occurs
violent interactions expected
(b) active
safe, no imminent danger
no violent interactions
preparation
PA is buying equipment at the start of a new round
inactive
PA is dead or game has not started yet
(a) passive
dead/no interaction

Figure 1. The four boxes explain the different semantic game classes used in this study and how they relate to the categories used by Mathiak and Weber [11], which are displayed in the dashed nested boxes. The classes are ordered from bottom to top in terms of increasing violent content, where PA stands for “player’s avatar”.

Category 3 was further divided into “search” and “potential danger”, and for category “violence” it is distinguished whether the PA is injured/attacked or fighting actively. However, automatic distinction of these semantic classes would not be feasible without neglecting the target to have a generic video content analysis system. For example, consider the highly abstract semantics regarding the distinction of “search” and “danger”. When the PA currently is in the state “search” (no imminent danger) and spots another character, its state switches to (potential) “danger”. Now, according to whether this character is identified as an enemy or not, the state switches to “violence”, because the PA shoots at the enemy, or back to “search” when the appearing character is harmless. Normally, state “danger” endures only for a few seconds before the states evolve further in the mentioned manner. Furthermore, the appearance of new characters in the PA’s field of view often takes place near the horizon, where avatars are only a few pixels in size, and it is extremely difficult to perform the necessary friend-or-foe identification with a reasonable precision. Furthermore, our system does not distinguish between “active” and “passive” violence. In practice, “passive” violence is a very short segment before either “active” violence or “inactive” (player’s avatar is dead) take place. This is the reason for the definition of the four classes described above: In this way, an automatic and generic annotation system is feasible and the remaining manual revisions are minimized. Figure 2 shows example frames for each of the four semantic game categories.

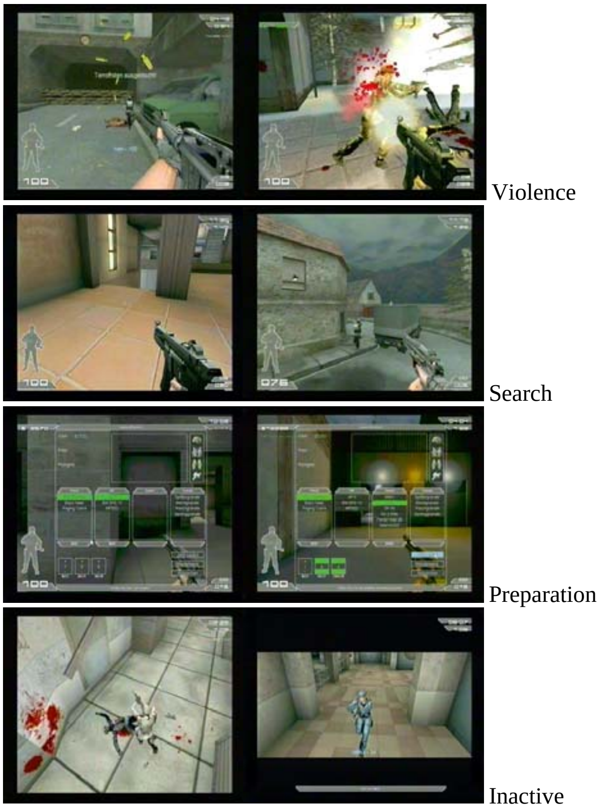

Figure 2. Example frames of semantic game categories.

3.2 Extraction of Audio Features

The semantic content of computer games is present in all modalities of their recordings: fighting and killing, for example, is visible in the video domain by the presence of enemies, muzzle flash and blood; it is also audible in the accompanying soundtrack by means of shots or explosive sounds as well as moans. The automatic content analysis system extracts a number of general audio low-level features which support the recognition of the semantic classes. The following features are extracted from non-overlapping 25ms frames [10] and are fed directly into the annotation system:

1. Eighth-order Mel Frequency Cepstrum (MFC) Coefficients:
   Capturing the broad envelope of the spectrum;
2. Zero Crossing Rate:
   A measure of oscillation and intra-frame variation;
3. Short Time Energy:
   Corresponding with loudness;

4. Sub-band Energy Distribution:
   Loudness ratio for four successive frequency bands;
5. Brightness and Bandwidth:
   The spectrum's frequency centroid and spread;
6. Spectrum Flux:
   Inter-frame spectral variation;
7. Band Periodicity
   Periodicity of the four subbands;
8. Noise Frame:
   Noisiness corresponding to lack of periodicity.

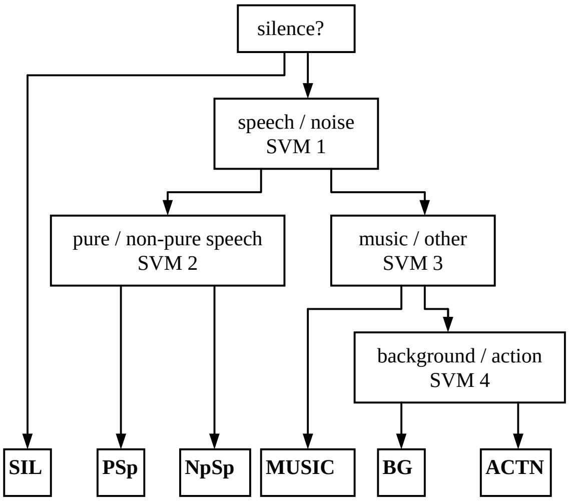

Figure 3: Scheme of the hierarchical audio type classifier: A single feature vector per sub-clip serves as input; output is a single acoustic class label and its corresponding probability.

Additionally, these features are fed into a content-based audio classification and segmentation system based on the approach of Lu et al. [10]. The system produces mid-level features on a per-second (sub-clip) basis in the form of acoustic class labels and related probabilities for silence (SIL), pure-/non-pure speech (PSp/NpSp), music, background (BG) and action sounds (ACTN). The low-level features are therefore aggregated per second, normalized and then concatenated to form one feature vector per sub-clip, which is processed by a hierarchical tree of SVMs, if it was not previously classified as silence by a threshold based classifier. Figure 3 shows this classification tree. It is trained on more than 32 hours of audio – TIMIT [9] data for clean speech, NOIZEUS [7] and broadcast speech data for non-pure speech, pop and instrumental music, various movie sound samples from broadcast material, and free web resources for the different types of noise. Five-fold cross-validation on a subset of 15000 feature vectors has been used to find the best parameter settings for a one-class SVM with RBF (radial basis function) kernel via libSVM [1]. The final acoustic class labels and their respective probabilities are fed into the game state learning algorithm as mid-level features to further guide the discovery of semantic patterns.

3.3 Extraction of Visual Features

Several visual features are extracted for each video frame. In addition to low-level features as color moments and texture features, several mid-level features are extracted automatically by utilizing camera motion estimation [4], face detection [13] and text detection [6]. In the following, the extracted features are briefly described:

- Color moments: Color moments are extracted at two different granularities. The first three global color moments are computed for the whole image. Corresponding values are extracted for each region of a 3 x 3 grid in HSV (Hue, Saturation, Value) color space. The i-th pixel of the j-th color channel of an image region is represented by c_{ij}. Then, the first three color moments are defined as:

$$mean_j = \frac{1}{N} \cdot \sum_{i=0}^{N-1} c_{ij} \tag{1}$$

$$stdev_j = \sqrt{\frac{1}{N} \cdot \sum_{i=0}^{N-1} (c_{ij} - mean_j)^2} \tag{2}$$

$$skew_j = \sqrt[3]{\frac{1}{N} \cdot \sum_{i=0}^{N-1} (c_{ij} - mean_j)^3} \tag{3}$$

- Texture features: The gray-scale image co-occurrence matrices m_k are constructed at 8 orientations. We use these matrices to extract the following values representing the global texture:

$$energy_k = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (m_{kij})^2 \tag{4}$$

$$contrast_k = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (i - j)^2 \cdot m_{kij} \tag{5}$$

$$entropy_k = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} m_{kij} \cdot log(m_{kij}) \tag{6}$$

$$homogeneity_k = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \frac{m_{kij}}{1 + |i - j|}, \tag{7}$$

  where N is the number of gray values and m_{kij} is the value of the co-occurrence matrix m_k at position (i, j).
- Camera motion features: Videos are segmented into shots using the cut detection approach described in [3]. Motion vectors embedded in MPEG videos are employed to compute camera motion at the granularity of P-frames, according to the approach presented in [4]. The following camera motion types are distinguished: translation along the x-axis, respectively y-axis, rotation around the x-axis, respectively y-axis and z-axis, and zoom.
- Text features: A robust text detection approach [6] is applied which can automatically detect horizontally aligned text with

different sizes, fonts, colors and languages. First, a wavelet transformation is applied to the image and the distribution of high-frequency wavelet coefficients is considered to statistically characterize text and non-text areas. Then, the k-means algorithm is used to classify text areas in the image. The detected text areas undergo a projection analysis in order to refine their localization. We use the detected text areas to derive the following features: number of text elements, distribution of text elements, and text frame coverage.

- Face features: Frontal faces are detected in each video frame using the face detector provided by Intel's OpenCV library [www.intel.com/technology/computing/opencv]. The number of detected faces and the face frame coverage are considered as feature values.

The camera motion features are useful to recognize the game state of searching and exploring, whereas text detection and texture features help recognizing the preparation state. A player steps into the preparation state with the intention to maintain his/her equipment. This screen contains several menus and is characterized by a high proportion of overlaid text. Thus, text features are assumed to be very good criteria to detect preparation states. However, text detection in the used game videos is a challenging task, because the text is printed on complex background and the frames include many MPEG artifacts. For the first-person-shooter game "Tactical Ops: Assault on Terror" color moment features seem to be useful to detect the state inactive because of the mostly appearing black areas at the top and bottom of the screen.

3.4 Semantic Classification

The goal of the proposed system is to learn models for the high-level semantic states of video games described in section 3.1 based on the extracted audiovisual low-level and mid-level features. As stated above, we do not focus on special properties of the computer game under consideration ("Tactical Ops: Assault on Terror"). Instead of using a specific and narrow approach that only works for a single video game, a generic video content analysis system is utilized that can be easily adapted to other games or video genres. We have used the SVM suggested in [14] with improvements of Keerthi et al. [8] to learn the mapping between the extracted audiovisual features and the semantic game classes. We employ multimodal analysis using an early fusion scheme. The datasets consequently consist of concatenated audio and visual features. The training of the SVM is realized by Sequential Minimal Optimization [14]. This is a fast training method which scales somewhere between linear and quadratic in the training set size. We have investigated several strategies to classify the computer game videos which are described below.

3.4.1 *Classification Using the Baseline System*

Several SVMs (one for each game class) must be combined to solve our problem, since SVMs are binary classifiers. To make a decision about the game state of a certain frame, the SVM models are employed to provide probability scores for a test instance (frame). These scores are compared and the class with the highest score is chosen.

3.4.2 *Classification Using Temporal Neighborhood*

It is observable that the appearance of a certain class is reflected also by the probability scores which are assigned to neighbored frames by an initial SVM classifier. This is the motivation for our second strategy to classify the computer game content. In addition to the audiovisual features, some time series information is utilized. The basic idea of this strategy is to obtain information about the temporal neighborhood of a frame using the probability scores of the initial SVM classifier. Based on the classification results, the relative frequency of each class in the temporal environment is computed for the current frame. The relative frequency of class c in the neighborhood of frame k is calculated according to the following formula:

$$freq_c(instance_k) = \frac{1}{2w+1} \cdot \sum_{i=k-w}^{k+w} t_c(instance_i)\,, \quad (8)$$

with $t_c(instance_i)$=1 if frame i is classified as class c and 0 otherwise, and w defines the window size. For example, if the relative frequency of violence is 0.5 for a frame, it follows that 50% of the neighboring frames are classified as violence. Furthermore, a smoothing filter is applied to the class probabilities obtained by the initial classifier. In both cases, a sliding window size of 25 frames is applied. We use the probability scores of the initial classifier, the frequencies and the smoothed values (4 features each) as new features and then re-train another classifier that makes the final decision. The processing steps are displayed in figure 4.

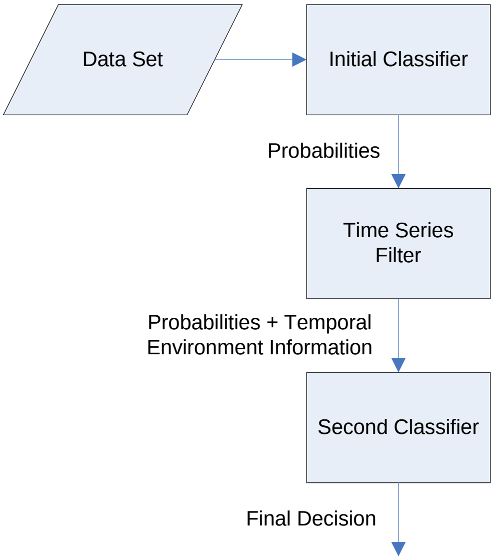

Figure 4. Concatenation of classifiers in order to employ temporal information.

3.4.3 *Refinement Using Semi-Supervised Learning*

In the setting of the addressed psychological experiment, the consumers always play the same game but at different levels and hence, they explore different virtual environments. Thus, it is possible that the SVM models learned from the training video are not suited well to distinguish between the different game classes in the test video. In previous papers [2][5], we have shown that an initial model obtained via unsupervised learning can be improved adaptively for a particular video. In order to achieve a more robust classification for a particular game video in our scenario, we employ a similar idea and propose a semi-supervised learning approach. A machine learning approach is called semi-supervised when unlabeled samples are incorporated in the training process. In our case, these are all frames of the test video since the class labels are unknown for them.

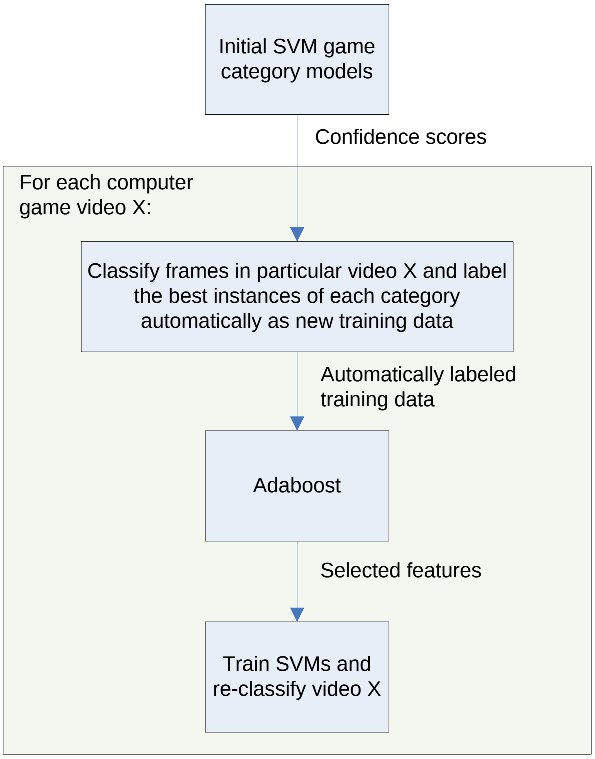
Initial SVM game category models
Confidence scores
For each computer game video X:
Classify frames in particular video X and label the best instances of each category automatically as new training data
Automatically labeled training data
Adaboost
Selected features
Train SVMs and re-classify video X

Figure 5. Main processing steps of the semi-supervised learning approach.

The main processing steps depicted in figure 5 are as follows. First, we use the training video to build a classifier consisting of the initial game category models. The initial classifier is used to classify the instances (frames) of a test game video as described in section 3.4.1. Then, the instances are ranked separately for each game category based on the probabilities of the detected classes. These rankings are then used to choose the instances with the highest confidence of each class. The top fifty percent of each class of the automatically labeled instances are chosen as positive training samples. Based on these automatically labeled most relevant instances of the test video, relevant features are selected using Adaboost [18]. The most relevant 77 features are chosen for subsequent use. An additional classifier is built using the previously chosen instances and selected features. Finally, this semi-supervised classifier, consisting of four newly trained SVMs, is used to classify the test video.

4. EXPERIMENTAL RESULTS

In this section, we present several experiments to test the system's applicability for the psychological study. The main goal is to significantly reduce the human annotation effort while achieving an accuracy that is comparable to a manual annotation. In the original experimental setting, the human annotators needed 120 hours to label the entire video collection [19]. In addition, the goal was to keep the video content analysis approach generic.

Four computer game videos were used to evaluate the system performance. All computer game videos show a resolution of 352 x 288 pixels and a video frame rate of 25 frames per second. Table 1 presents the distribution of the semantic game categories for each of the used videos. The ground truth data were created by Weber et al. as described in [19].

Table 1. Number of frames referring to semantic game categories for each of the used computer game videos.

	Preparation	*Search*	*Violence*	*Inactive*	*Total*
Game-vmj3_7	2390	11657	488	2155	16690
Game-vmj6_3	1665	8574	525	5601	16365
Game-vmj6_4	2364	6445	2630	5251	16690
Game-vmj6_5	2157	10023	1211	2581	15972

A "leave k-1 videos out" cross validation scheme is used: Since the main goal is the reduction of human annotation effort, only one video is used as training data in each test while the remaining three videos are used as test videos. The SVM has been implemented using the WEKA library [20]. A radial basis function kernel was used for the SVM. Adaboost has been implemented according to the description given in [18].

The following system variations were tested: 1.) The first one is the baseline system as described in 3.4.1. All features mentioned in section 3.2 and 3.3 are used to learn a SVM model for each semantic game class; 2.) After an initial SVM training, further features are generated that capture temporal characteristics of classes as described in section 3.4.2; 3.) The semi-supervised learning scheme as described in section 3.4.3: after an initial classification of a test video, the frames that are classified with highest confidence are used as training data. These training data are used to learn new SVM models, and finally the same video is classified using these models.

The results for these experiments are presented in Table 2 - Table 4. The following definitions are used to evaluate the results:

$$recall = \frac{\#\, correctDetectedItems}{\#\, Items} \qquad (9)$$

$$precision = \frac{\#\, correctDetectedItems}{\#\, correctDetectedItems + \#\, falseAlarms} \qquad (10)$$

$$f1 = \frac{2 * recall * precision}{recall + precision} \qquad (11)$$

Table 2. "Baseline" system: Recall, precision and f1-measure for each of the four semantic classes as well as the total recall.

	Prepa-	*Search*	*Vio-*	*In-*	*Total*
	ration		*lence*	*active*	*recall*
Recall	84.3	92.3	53.9	88.5	87.5
Precision	86.0	87.5	68.7	93.4	
F1	85.1	89.8	60.4	90.9	

Table 3. "Baseline + temporal features": Recall, precision and f1-measure for each of the four semantic classes as well as the total recall.

[%]	*Prepa-*	*Search*	*Vio-*	*In-*	*Total*
	ration		*lence*	*active*	*recall*
Recall	83.1	92.6	56.7	91.6	88.5
Precision	87.7	88.5	68.4	94.1	
F1	85.4	90.5	**62.0**	92.8	

Table 4. "Baseline + Semi-Supervised Learning": Recall, precision and f1-measure for each of the four semantic classes as well as the total recall.

[%]	*Prepa-*	*Search*	*Vio-*	*In-*	*Total*
	ration		*lence*	*active*	*recall*
Recall	92.2	94.9	55.3	92.0	**91.0**
Precision	96.0	90.0	66.0	97.6	
F1	**94.1**	**92.4**	60.2	**94.8**	

Several observations can be made. At first, our automatic baseline system achieves a frame-based total recall of 87.5% on the average. This is a very good result if one considers that the inter-coder reliability in the original psychological experimental setting between the human annotators was 0.85 [19]. In nearly any experiment, preparation, search and inactive states were recognized well, whereas the recognition of violent states is rather difficult. In terms of total recall, the semi-supervised approach outperforms the alternative approaches (see Table 5). The approach using temporal neighborhood information achieves the best performance for the most difficult concept "violence" and recognizes more than half of the violent actions correctly while keeping the precision at nearly 70% at the same time. The confusion matrix in Table 6 allows gaining insight in the system failures of the best system (semi-supervised learning). The diagonal represents the number of correctly classified frames. For example, the most frequent error is that a violence frame is misclassified as search, and vice versa, whereas e.g. a violence frame was never classified as preparation. Overall, we conclude that our proposed system achieves a very satisfying performance. It demonstrates the ability to reduce human annotation efforts to a minimum because the system automatically determines relevant game events with high reliability.

Table 5. Total recall for each of the tested systems.

[%]	*Baseline*	*Tem-*	*Semi-Sup-*
		poral	*Learning*
Total recall	87.5	88.5	91.0

Table 6. Confusion matrix for the semi-supervised learning experiment. For example, the most frequent error is that a violence frame is misclassified as search and vice versa.

			Vio-	*In-*
	Prep.	*Search*	*lence*	*active*
	(GT)	*(GT)*	*(GT)*	*(GT)*
Det. Prep.	23733	918	52	8
Det. Search	1991	104535	6449	3194
Det. Violence	0	3623	8046	520
Det. Inactive	4	1021	15	43042

5. CONCLUSIONS

In this paper, we have presented an automatic semi-supervised semantic video analysis system that supports psychological experiments with respect to violence in computer games. In the addressed interdisciplinary study, annotations are required to find interrelationships between the consumer's brain activity and game events during the recorded game sessions, in particular with respect to violent actions. Our proposed system automatically labels such videos and achieves a total recall of up to 91% in the best case using a semi-supervised learning approach. This approach adaptively refines a model on a particular video: Based on the initial classification result, the approach automatically labels the frames in a new video and adapts its concept models to this video by employing feature selection to adaptively learn a classifier for a particular game video.

Considering the fact that Weber et al. [19] observed an inter-coder reliability of 0.85 for human annotators, our automatic system demonstrates an excellent performance. In addition, since our semi-supervised approach needs labeled training data for a single video only, the required human supervision could be kept at a minimum in this interdisciplinary study. The graphical user interface of our software Videana enables a human expert to refine respectively correct the annotation results: As a basic requirement, the annotations must be as accurate as possible to investigate the interrelationship with a player's brain activity. However, such a correction step must also be applied when only human annotators label the videos. Overall, we conclude that the experimental results demonstrate the applicability of our system for the interdisciplinary studies in the field of media and behavioral sciences.

There are several areas for future work. For example, semantic concepts like "danger" in our scenario are very difficult to recognize. This concept depends on the detection of person occurrences and particularly on a "friend-or-foe" distinction. However, these persons appear in very small sizes in the game, and it is even hard for a human annotator to make a decision whether the situation is actually “dangerous” or not. Finally, temporal state transitions promise to entail additional useful information, e.g. the state “inactive” is always preceded by the state “violence”. Such temporal relationships should also be incorporated in the automatic annotation system.

6. ACKNOWLEDGMENTS

This work is financially supported by the Deutsche Forschungsgemeinschaft (SFB/FK 615, Project MT).

7. REFERENCES

[1] Chang, C. and Lin, C. libSVM: A Library for Support Vector Machines, 2001, available online (09.02.2007): http://www.csie.ntu.edu.tw/~cjlin/libsvm

[2] Ewerth, R. and Freisleben, B. Self-Supervised Learning for Robust Video Indexing. In *Proceedings of the IEEE Conference on Multimedia & Expo*, Toronto, 2006, pp. 1749-1752.

[3] Ewerth, R. and Freisleben, B. Video Cut Detection without Thresholds. In *Proceedings of the 11th International Workshop on Systems, Signals and Image Processing*, Poznan, Poland, 2004, pp. 227-230.

[4] Ewerth, R., Schwalb, M., Tessmann, P., and Freisleben, B. Estimation of Arbitrary Camera Motion in MPEG Videos. In *Proceedings of the 17th International Conference on Pattern Recognition*, Vol. 1, Cambridge, United Kingdom, 2004, pp. 512-515.

[5] Ewerth, R., Mühling, M., and Freisleben B. Self-Supervised Learning of Face Appearances in TV Casts and Movies. In *Proceedings of the 8th IEEE International Symposium on Multimedia*, San Diego, USA, 2006, pp. 78-85.

[6] Gllavata J., Ewerth R., and Freisleben B. Text Detection in Images Based on Unsupervised Classification of High-Frequency Wavelet Coefficients. In *Proceedings of 17th International Conference on Pattern Recognition*, Vol. 1, Cambridge, UK, 2004, pp. 425-428.

[7] Hu, Y. and Loizou, P. Subjective Comparison of Speech Enhancement Algorithms. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, Toulouse, France, 2006, pp. 153-156.

[8] Keerthi S., Shevade S., Bhattacharyya C., and Murthy K. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*, Vol. 13, 2001, pp 637-649.

[9] Linguistic Data Consortium, The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, 1990, available online (09.02.2007): http://www.ldc.upenn.edu/Catalog/readme_files/timit.readme.html

[10] Lu, L., Zhang, H., and Li, S. Content-based Audio Classification and Segmentation by Using Support Vector Machines. *Multimedia Systems*, Vol. 8, Springer, 2003, pp. 482-492.

[11] Mathiak, K. and Weber, R. Towards Brain Correlates of Natural Behavior: fMRI During Violent Video Games. *Human Brain Mapping*, Vol. 27, 2006, pp. 957-962.

[12] Naphade, M. and Smith J. On the Detection of Semantic Concepts at TRECVID. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, New York, USA, ACM Press, 2004, pp. 660-667.

[13] OpenCV, Intel Open Source Computer Vision Library, available online (09.02.2007): http://www.intel.com/technology/computing/opencv/

[14] Platt, J. Fast Training of Support Vector Machines using Sequential Minimal Optimization. *Advances in Kernel Methods - Support Vector Learning*, MIT Press, 1999, pp. 185-208.

[15] Sadlier, D. and O’Connor, N. Event Detection in Field Sports Video using Audio-visual Features and a Support Vector Machine, In *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 15 (10), 2005, pp. 1225-1233.

[16] Tong, X., Liu Q., Duan, L., Lu, H., Xu C., and Tian, Q. A Unified Framework for Semantic Shot Representation of Sports Video, In *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, Hilton, Singapore, 2005, pp. 127-134.

[17] TRECVID: TREC Video Retrieval Evaluation Homepage, available online (09.02.2007): http://www-nlpir.nist.gov/projects/trecvid/

[18] Viola, P. and Jones, M. Robust Real-Time Face Detection. In *International Journal of Computer Vision*, Vol. 57 (2), Kluwer Academic Publishers, 2004, pp. 137-154.

[19] Weber, R., Ritterfeld, U., and Mathiak, K. Does Playing Violent Video Games Induce Aggression? Empirical Evidence of a Functional Magnetic Resonance Imaging Study. *Media Psychology*, Vol. 8, 2006, pp. 39-60.

[20] Witten I. and Frank E. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[21] Xu, H. and Chua, T. Fusion of AV Features and External Information Sources for Event Detection in Team Sports Video. In *ACM Transactions on Multimedia Computing, Communications, and Applications*, Vol. 2 (1), 2006, pp. 44-67.