VIOLENT SCENES DETECTION USING MID-LEVEL VIOLENCE CLUSTERING

Shinichi Goto[1] and Terumasa Aoki[1,2]

[1]Graduate School of Information Sciences,Tohoku University, Miyagi, Japan
s-goto@riec.tohoku.ac.jp
[2]New Industry Creation Hatchery Center,Tohoku University, Miyagi, Japan
aoki@riec.tohoku.ac.jp

ABSTRACT

This work proposes a novel system for Violent Scenes Detection, which is based on the combination of visual and audio features with machine learning at segment-level. Multiple Kernel Learning is applied so that multimodality of videos can be maximized. In particular, Mid-level Violence Clustering is proposed in order for mid-level concepts to be implicitly learned, without using manually tagged annotations. Finally a violence-score for each shot is calculated. The whole system is trained ona dataset from MediaEval 2013 Affect Task and evaluated by its official metric. The obtained results outperformed its best score.

KEYWORDS

Multimedia Analysis, Video Processing, Machine Learning

1. INTRODUCTION

The amount of videos on the Internet has been rapidly increasing in recent years, which enables people to access them easily, and has given their lives entertainment. This situation also makes it possible for children to easily reach violent contents at the same time though it should be filtered. Because of the enormous number of them, however, it is almost impossible to give annotations on those videos manually to remove them. This makes it essential to develop the automatic classification system for violent videos. As a matter of fact, Technicolor [30] is proposing the need of a system that enables users to choose movies that are suitable for children in their families by providing a preview of violent segments beforehand in MediaEval [1]. MediaEval is a benchmarking workshop dedicated to evaluating algorithms for multimedia analysis and retrieval. They have started arranging Affect Task, which is intended to detect violent scenes in movies.

In spite of this situation, Violent Scenes Detection still has much difficulty because of its complexity, as well as its ambiguous definition. For instance, Chen et al. defines violence as "a series of human actions accompanying with bleeding" in [2], though Giannakopouloset al. only defines violent-related classes such as shots, fights and screams in [3].Or some papers have no enough description for the dataset used in detail according to the research in [4].

In this paper, we propose a novel system to detect violent scenes in movies, using a violent definition by MediaEval 2013 Affect Task, which is "*physical violence accident resulting in human injury or pain*." Our system is based on segment-level processing. First movies are

David C. Wyld et al. (Eds) : CCSIT, SIPP, AISC, PDCTA, NLP - 2014
pp. 283–296, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4224

separated to segments, each of which has a fixed number of frames, and both of visual and audio feature vectors for each segment are extracted. Those feature vectors are used to train classifiers. In order to make the most use of multimodality of movies, Multiple Kernel Learning is applied for our system. In addition, Mid-level Violence Clustering is proposed in order for implicit violent concepts to be learned, without using annotations tagged manually by humans. Classifiers produce segment-level violence-scores, and finally they are converted to shot-level scores. Our system is trained and tested on a dataset from MediaEval 2013 Affect Task, and evaluated by its official metric MAP@100. We compare our results with results by participants in MediaEval. Moreover, an investigation for each mid-level violence cluster is performed for further understanding.

2. RELATED WORK

Compared with other related researches on video analysis such as *event classification* or *action recognition*, relatively few researches have been done for Violent Scenes Detection. Giannakopouloset al. [3] analyzed the effect of audio features on this task, such as energy entropy, zero crossing and so on. For each segment, those audio features are extracted, and fed as input for Support Vector Machine(SVM) without visual features. Bermejo et al., on the other hand, used only visual features [5]. Features such as Bag-of-Visual-Words (BoVW) [14], Space Time Interest Points (STIP)[21] and MoSIFT [22] are calculated for SVM. For a SVM kernel function, they compared Radial Basis Function (RBF), Chi-square and Histogram Intersection Kernel (HIK) [17], and reported that HIK obtained the best result. Chen et al. also used only visual information such as average motion vectors, camera motion, shot length and RGB values of pixels. They also proposed a shot-grouping algorithm for the further efficiency [2].

Though researches above use either visual or audio features, researches utilizing both of them have shown to improve results. The first try that utilized this multimodality is a work by Nam et al. at 1998 [6]. They characterized and indexed violence scenes in videos, proposing that violent signatures are represented as combination of multiple features. Their feature extraction is based on flame detection, blood detection and audio features such as energy entropy. Lin et al. [7] adopted PLSA to locate audio violence. PLSA is a probabilistic model utilizing the Expectation Maximization algorithm, which is often used in the field of natural language processing. For visual violence they used a linear weighted model fed with the results of violent event detections such as motion intensity, frame, explosion and blood. Finally two classifiers are built in a co-training way. Penet et al. also utilized both modalities, comparing normal Early Fusion with Late Fusion [8]. Early Fusion concatenates features from both modalities before learning, while Late Fusion fuses the probabilities of both modalities. They reported that Late Fusion has more effectiveness. Derbas et al. [20] proposed Joint Audio-Visual Words representation, which constructs codebooks in the context of Bag-of-Words (BoW)by combining audio and visual features. They got the first in MediaEval 2013 without external data in terms of MAP@100 score. Instead of detecting violent scenes directly from low-level features, some works have used mid-level concepts such as *Fire*, *Fights* or *Explosions* given in MediaEval Affect Task [1]. Ionescu et al. proposed a frame-level violence prediction, applying a multi-layer perceptron in order to utilize these concepts [9][25].They put the first layer for the concept prediction, and the second layer for the violence prediction. They ranked first in MediaEval 2012 onF_1-$score$. In addition to those given concepts, Tanand Ngo [10]have utilized extra 42 violence concepts such as *bomb* or *war* from ConceptNet [23], which is composed of nodes representing concepts in the form of words or short phrases with their relationships. Their system trains those extra concepts using YouTube videos which they crawl additionally. Afterwards a graphical model of those concepts are generated, and Conditional Random Fields [31]refines it by using relationships in ConceptNet and co-occurrence information of concepts. Their MAP@100 result was the first place in 2013 Affect Task with external data.

3. VIOLENT SCENES DETECTION BASED ON MID-LEVEL CLUSTERING

3.1. Approach Overview

The overview of our system, which is composed of the training process and the testing process, is displayed in Figure 1. For both processes, feature extraction and training/classification are based on the segment-level calculation. Here we define a *segment* as a bunch of 20 frames(0.8 seconds if FPS is 25).

First all training movies are separated to segments. Then both of visual and audio features are extracted for each segment (described in 3.2). Segments that are tagged as violent are gathered and separated $K(> 0)$ clusters, each of which represents a concept or a combination of concepts that is led to violence in our assumption. This will be described in more detail in 3.3. For each cluster a classifier is trained by using Multiple Kernel Learning (MKL).

In the testing process, segmentation andfeature extraction are performed in the same way as the training process.Classifiers in all clusters evaluate each segment, producing violence-scores. Then scores are integrated to generate segment-level scores.Smoothing is applied in order to take the context of movies into account, and finally segment-level scores are converted to shot-level scores.

For our runs only violent and non-violent ground truth are used, and neither a high-level concept nor external data is used. The following sections explain our feature vectors and training system more precisely.

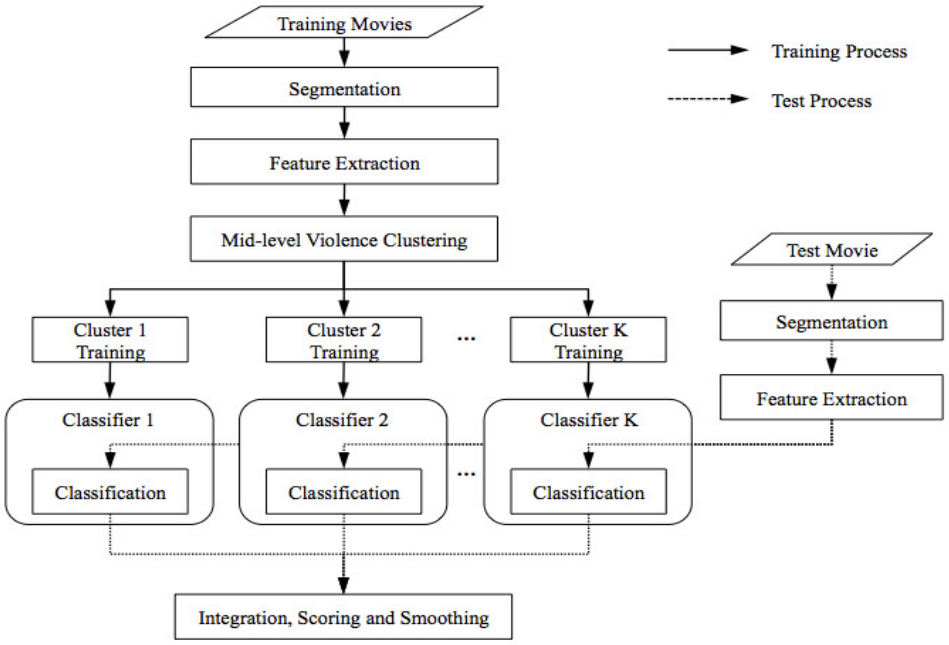

Figure 1. Approach Overview.

3.2. Features

Recent works on violent scenes detection such as [10] and [24]have shown the effectiveness of using trajectory-based features as visual information, MFCC-based features as audio information. Similar to those researches, in total six feature spaces exist on our system. Trajectory, HOG, MBHx, MBHy and RGB histogram around trajectories are computed as visual features, and MFCC and delta-MFCC are calculated as audio features. For each segment these features are converted to the BoW form by using already generated codebooks. Codebooks are calculated by using randomly selected 100,000 features and k-means++ algorithm [33] before hand in each feature space respectively.

3.2.1. Visual Features

Trajectories, which can capture local motion of videos, are getting attentions especially in the field of *action recognition*. Wang et al. applied dense sampling, which is used in the field of image classification, to trajectories to improve their quality, and called them Dense Trajectories [11]. Except for those in homogeneous areas represented by a small value of the eigenvalue of the auto-correlation matrix, densely sampled feature points are tracked by calculating optical flows in each spatial scale until they reach the length of $L = 15$frames. Every frame newly sampled points are added if no tracked point is found in the neighborhood of each pixel.

For each trajectory, descriptors are extracted in its neighborhood. In addition to existing descriptors extracted along trajectories such as Histograms of Oriented Gradients (HOG) or Histograms of Oriented Optical Flow (HOF), they also proposed to extract Motion Boundary Histograms (MBH), which was originally proposed in the field of human detection by Dalal et al. [12]. MBH represents the changes in the optical field, namely local motion information independent of camera motion by calculating the gradient of the optical flow. MBH is calculated separately along vertical direction (MBHx) and horizontal direction (MBHy).

Descriptors are calculated in 12(=2*2*3) subdivided volumes around trajectories and concatenated afterwards. For our system, HOG (96-dimension), MBHx (96-dimension) and MBHy (96-dimension) around trajectories are extracted. Also displacement vectors of trajectories are extracted for both of x-direction and y-direction (30-dimension). Because of the frequent camera motion, HOF is expected to have poor contribution on our task, and therefore is removed. On the other hand, as color information is expected to be helpful just as blood or flame detections contributed to the results in some previous researches, 64-bin RGB histograms around trajectories are also calculated every 5 frame, which will be 192-dimension descriptor. Finally they are converted to the form of BoW for each segment, and 200-dimension trajectory, 400-dimension HOG, 200-dimension MBHx, 200-dimension MBHy, and 400-dimension RGB histogram are obtained. As parameters we used 32 for a neighbour range, 5 for a sampling step, and 6 for a spatial scale size.

3.2.2. Audio Features

Similar to Bag-of-Audio-Words in [13], MFCC (Mel Frequency Cepstrum Coefficients) and the log energy are first extracted every 10ms with 5ms overlap for audio features. The first derivative of MFCC and its energy are also calculated as delta-MFCC, which results in 26-d features in total. Then they are converted to 200-d BoW for each segment.

3.3. Mid-level Violence Clustering

Before training classifiers by using features above, Mid-level Violence Clustering is applied. Figure 2 illustrates the construction process of mid-level clusters. Note substitutefigures are displayed for easy understanding instead of feature vectors.First whole violent segments in training movies are gathered. Then they are separated to K (> 0) clusters, each of which has similar segments. Then non-violent training segments are assigned to one of those clusters randomly, whose results construct clusters for mid-level violence classifiers. Practically we

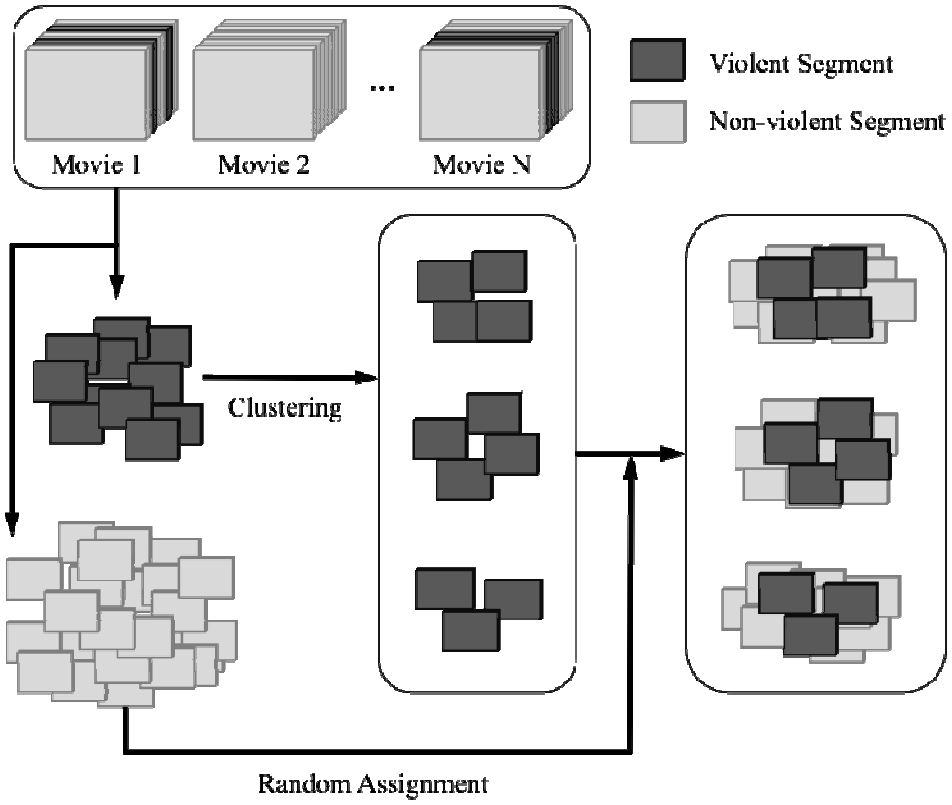

Figure 2. Mid-level Violence Clustering ($K = 3$).

need to handle 6 feature vectors while clustering. We simply concatenates them, and applied k-means++ algorithm with Euclidean distance. This means Euclidian distance is used for the *similarity* among segments. Here we assume that in each cluster feature vectors for violent segments represent multiple mid-level concepts related to violence.

The reason why we design this clustering process is that the diversity of "violence" is huge: even if two segments are tagged as violent, their features might be largely different depending on their characteristics of "violence."For instance, although explosion scenes labelled as violent might have distinctive visual features, those of scream scenes might not similar even if they are also labelled as violent. A schematic example simplifies this problem in 2-dimension (see Figure 3a). There exist two classes (Class A and Class B), and what we want to do is to define a decision boundary that can classify new input points. As Figure 3a illustrates, data is noisy and this classification problem does not seem to be easily solved. Here let us suppose that Class A has two mid-level concepts, and those concepts can be separated to two clusters (Figure 3b and Figure 3c).In that case training and classifying Class A in each figure corresponds to training and

classifying each mid-level concept in Class A, which is much easier problem than directly dealing with Figure 3a.

In violent scenes detection, as a few previous works such as [9] and [10] have shownmid-level concepts information can be helpful, it is appropriated to say that "violence" class has multiple mid-level concepts. Hence if those concepts are correctly clustered, training violence for one cluster corresponds to training one mid-level violent concept. In addition, being different from the problem in Figure 3, the number of feature points and dimensions are huge on our task.As 10 concepts are prepared in MediaEval, and as in [10] 52 concepts have been used, the number of mid-level concepts wouldbe much bigger, which is led to more complexity. Thus

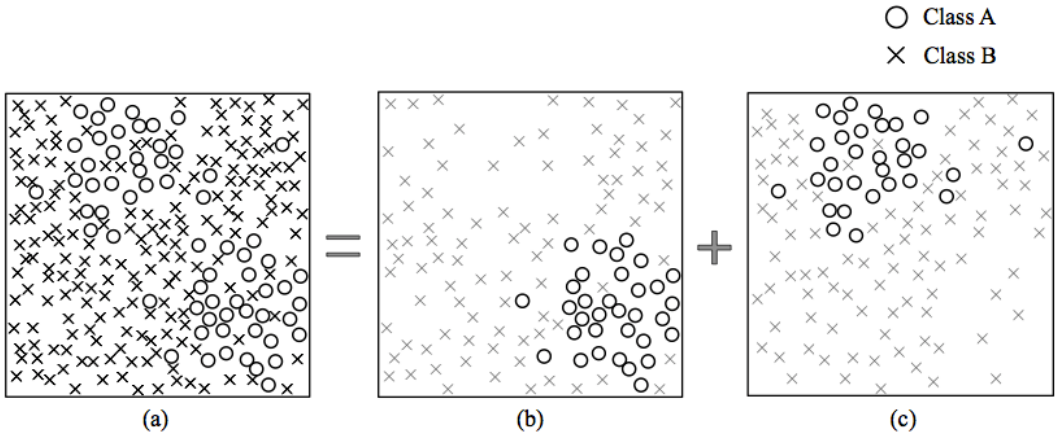

Figure 3. Example problem where mid-level clustering can be used ($K = 2$): (a) the original problem, (b) first cluster, (c) second cluster. We assume two clusters exist in (a).

simplifying this problem is expected to be effective. The process of actual training, classification and integration are described in the following sections.

3.4. Multiple Kernel Learning

Support Vector Machine (SVM) has had a great contribution to the field of image classification over the last few years. In particular, BoVW with SVM [14] had succeeded extremely. In violence detection in movies, however, multiple feature spaces have to be handled, and then simply concatenating feature vectors and training classifiers, which is called Early Fusion, is not always a best way, according to the work by Cedric et al. [8].As they studied, when multimodal features exist, there are two available fusion schemes: namely Early Fusion and Late Fusion. Early Fusion concatenates features from both modalities before training, which means it can take the correlation of those feature spaces into account while training, though it deals with each feature space uniformly. On the other hand, on Late Fusion training itself is done in each feature space independently, and extracted probabilities from the testing process are fused afterwards. They compared Early Fusion with Late Fusion process when two modalities exist (visual and audio).Although they concluded that Late Fusion has more effectiveness, if more modalities exist just as our system, it has some drawbacks: 1) it cannot take correlations of multimodal features into account while training, 2) how to fuse probabilities needs to be decided manually beforehand.

To cope with this problem, and to maximize the multimodality of our features, we apply Multiple Kernel Learning (MKL), which can be regarded as a kind of Early Fusion, but aims at finding optimized weights for each feature space when multiple SVM kernels are applied [15]. This means MKL considers correlations of multiple feature spaces, but at the same time it considers differences of violent characteristics among multiple feature spaces. Fusing method does not have

to be decided either. In MKL the whole kernel is composed of multiple sub-kernels, and is defined as the following equation:

$$\boldsymbol{K}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \sum_p \beta_p \boldsymbol{K}_p(\boldsymbol{x_i}, \boldsymbol{x_j}) \quad (1)$$

where $\boldsymbol{K}_p$ are sub-kernels, and $\beta_p$is a weight for p-thsub-kernel. In our case, kernels for Trajectory, HOG, MBHx, MBHy, RGB-histogram and Audio features are prepared.

When$N$data points $(\boldsymbol{x}_i, y_i)(y_i \in \{\pm 1\}, i = [1, 2, \cdots, N])$are given and input features$\boldsymbol{x}_i$ is translated via $\Phi_p(\boldsymbol{x}) \mapsto \mathbb{R}^{D_p}, p = [1, 2, \cdots, P]$into Pfeature spaces where D_p denotes the dimensionality of p-th feature space,Bach et al. derived the dual for the MKL primary problem in [16], which represents its optimization problem as following:

$$\begin{aligned} \min \quad & \boldsymbol{\gamma} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \alpha_i \quad \textbf{(2)} \\ \text{w.r.t.} \quad & \boldsymbol{\gamma} \in \mathbb{R}, \alpha \in \mathbb{R}^{\boldsymbol{N}} \\ \text{s.t.} \quad & 0 \le \alpha \le 1C, \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \alpha y_i = 0 \\ & \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{k}_p(\boldsymbol{x_i}, \boldsymbol{x_j}) \le \gamma, \forall p = 1, \cdots, P \end{aligned}$$

where$C$ is a regularization parameter and we have one quadratic constraint per kernel.

For a sub-kernel, Histogram Intersection Kernel (HIK), which has been reported to perform well on histogram-based features [17],is adopted. HIK is defined as follows:

$$\boldsymbol{K}_{int}(\boldsymbol{A}, \boldsymbol{B}) = \sum_{i=1}^{m} min(a_{i,} b_{i,}) \quad (3)$$

where $\boldsymbol{A} = [a_1, a_2, \cdots, a_m]and\boldsymbol{B} = [b_1, b_2, \cdots, b_m]$. It measures the degree of similarity between two histograms.

MKL is applied to all$K$ clusters, resulting in generating K classifiers.SHOGUN Toolbox [18] is used for our MKL implementation.

3.5. Integration, Scoring and Smoothing

After MKL, each segment in test movies is classified as violent or non-violent by K classifiers. Suppose a movie has N segments. For k-th classifier ($1 \le k \le K$), classification results for all segments are obtained:

$$\boldsymbol{C}_k = \left[c_{k,1}, c_{k,2}, \cdots, c_{k,N}\right] \ (c_{k,n} \in \{\pm 1\}, 1 \le n \le N) \quad (4)$$

where "$+1$" represents violence, while "-1" represents non-violence.

Let $d_{k,n}$denote a"violence-score," which is a distance between a feature point of n-thsegment and a hyperplane of k-th classifier.$\boldsymbol{D}_k(= \left[d_{k,1}, d_{k,2}, \cdots, d_{k,N}\right])$are calculated as the result of MKL, and we define$\boldsymbol{S}_k$, scores by k-thclassifier as follows:

$$\boldsymbol{S}_k = [s_{k,1}, s_{k,2}, \cdots, s_{k,N}], \qquad s_{k,n} = \begin{cases} d_{k,n}\ (if\ c_{k,n} = +1) \\ 0\ (if\ c_{k,n} = -1) \end{cases} \tag{5}$$

Scores by each classifier are integrated to produce pre-final scores $\boldsymbol{S}$:

$$\boldsymbol{S} = [s_1, s_{2,} \cdots, s_N], \ \ s_n = \frac{\sum_{l=1}^{K} s_{l,n}}{K_{vio}}\ (1 \le n \le N) \tag{6}$$

where$K_{vio}$is the number of classifiers whose c_k is "+1,"that is, the number of classifiers which classify a target segment as violent.This means for each segment, if no cluster classifies it as violent, its violence-score is zero, while the mean value of violence-scores of classifiers which classify it as violent is assigned if Nclusters classify it as violent.

So far the context of movies are not taken into account. Hence scores are smoothed as a final step. Although in [24] the average value over a three-shot window is calculated, we adopt a moving average calculation so that the further neighbour segments are positioned, the fewer their effects are considered. Smoothed scores $\acute{\boldsymbol{S}}$are calculated by using pre-final scores $\boldsymbol{S}$ as follows:

$$\acute{\boldsymbol{S}} = [\acute{s}_1, \acute{s}_{2,} \cdots, \acute{s}_M], \quad \acute{s}_i = \frac{s_i + \sum_{m=1}^{M} \alpha^m \cdot (s_{i-m} + s_{i+m})}{2M + 1} (0 < \alpha < 1) \tag{7}$$

where α is a smoothing coefficient, $M$is a neighbor range around a segment. We used 0.5 for α and 2 for M.

Scores for shots are calculated by converting segment-level scores after calculating frame-level scores. Because the numbers of frames in segments are consistent except for the final segment of the movie, frame-level scores are simply given as scores for segments which have those frames. Then for each shot scores for frames it has are summed and divided by the number of frames. This score is used as the final violence score for each shot. If this score is higher than a threshold, that shot is classified as violent. We choose 0.20 as a threshold.

4. EXPERIMENT

Our experiments are carried out following MediaEval 2013 Affect Task. Though in 2013 there are two subtasks and participants are allowed to submit multiple types of runs, in this paper we focus on shot-level classification with objective violence definition, which is "*physical violence accident resulting in human injury or pain*."Although there were participants who used additional data for their systems such as [10], our results are compared with systems without external data.

4.1. Dataset

Twenty-five movies are provided with shot boundary annotations.18 are dedicated to the training process: *Armageddon, Billy Elliot, Eragon, Harry Potter 5, I am Legend, Leon, Midnight Express, Pirates of the Caribbean 1, reservoir Dogs, Saving Private Ryan, The Sixth Sense, The Wicker Man, Kill Bill 1, The Bourne Identity, The Wizard of Oz, Dead Poets Society, Fight Club and Independence Day,* and they are given with frame-level violence ground truth. 7 are dedicated to the test process: *Fantastic Four, Fargo, Forrest Gump, Legally Blond, Pulp Fiction, The God Father 1 and The Pianist*. Though in MediaEval 2013 participants were allowed to use prepared high-level concepts, our algorithm use only low-level features extracted from movies and shot boundary information.

4.2. Evaluation Criteria

The official metric for the evaluation is the Mean Average Precision [19] at the 100 top ranked violent shots (MAP@100). Namely:

$$MAP@100 = \frac{1}{100}\sum_{j=1}^{100}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk}) \quad (8)$$

where $R_{jk}$is the set of ranked list of violent shots from $1st$ to k-thfor the calculation of Average Precision of j-th shot.

4.3. Overall Performance Evaluation

Table 1 presents our results with results by participants in MediaEval 2013.For the number of clusters Kwe tried 10, 30, 50 and 100.Additionally a score when we construct clusters for each training movie is displayed. This is the case in which violence multiple mid-level concepts are mixed in each cluster. We call this as Training-movie Clustering in this table.

A result with$K = 50$shows the best score 0.558, and one can find Mid-level Violence Clustering scores outperform the best score0.520 in MediaEval 2013. Also they are much higher than a score by training-movie clustering. As each cluster has multiple violent concepts

Table 1. Comparison between our results and results by participants in MediaEval 2013.

Team (Run)	MAP@100
LIG [20]	0.520
FAR [25]	0.496
Fudan [24]	0.492
NII [28]	About 0.400
Technicolor [29]	0.338
VISILAB [26]	0.150
MTM [27]	0.070
Mid-level Violence Clustering ($K = 30$)	0.527
Mid-level Violence Clustering ($K = 40$)	0.539
Mid-level Violence Clustering ($K = 50$)	**0.558**
Mid-level Violence Clustering ($K = 100$)	0.551
Training-movie Clustering	0.487

that might be largely different in Training-movie Clustering, this result supports our assumption that each cluster is supposed to have similar concepts in Mid-level Violence Clustering.

While a result seems to change a little depending on the number of clusters, even the high number of clusters (e.g. $K = 100$) can keep a promising score. This is because when the number of clusters is big, some clusters have smaller violent segments assigned to themselves due to their anomalies. For instance, when we chose $K = 100$, the smallest number of violent segments in a cluster was 2, although other clusters tended to contain 50-150violent segments. Even though this cluster could not find violent segments, it did not affect a final score either because our scoring equation (6) depends only on scores by classifiers that classify a target segment as violent.

Though our system achieves promising results, they are not high enough yet. The first point to be considered here is that our feature vectors might not be distinct enough. Although we have used only Trajectory-based features for visual information, they can be easily affected by camera

motion. Even though features such as MBH, which are proposed as robust to camera motion is extracted, they might be noisy if trajectories themselves are unreliable. Therefore some action against this problem is imperative.

The second reason is that thoughEuclidean distance is used for the *similarity* while Mid-level Violence Clustering, Histogram Intersection is used for sub-kernels in MKL, which might be inconsistent. To tackle this problem, two solutions can be considered: to perform clustering by using Histogram Intersection, or to apply sparse coding and max pooling [32] to whole feature vectors in order that linear kernel can be used for SVM.

Moreover, from our investigation, shots that contain frequent camera motion, multiple people and big sound tend to be miss-classified as violent. Loud sound such as abrupt brake by cars and crackers were also miss-classified. Meanwhile, common missed violent shots are violent scenes without sound, such as a scene in which a man is wringing other man's neck. Feature vectors that can detect these events are needed. Also in our system shot boundaries are not taken into account while constructing shots. Because shots often change in the middle of segments, feature vectors in those segments should be discarded.

4.4. Discussion

We still do not know how segments are classified by each cluster, or what kind of mid-level concepts is assigned to each cluster. Therefore we analyzed each cluster for its implicit

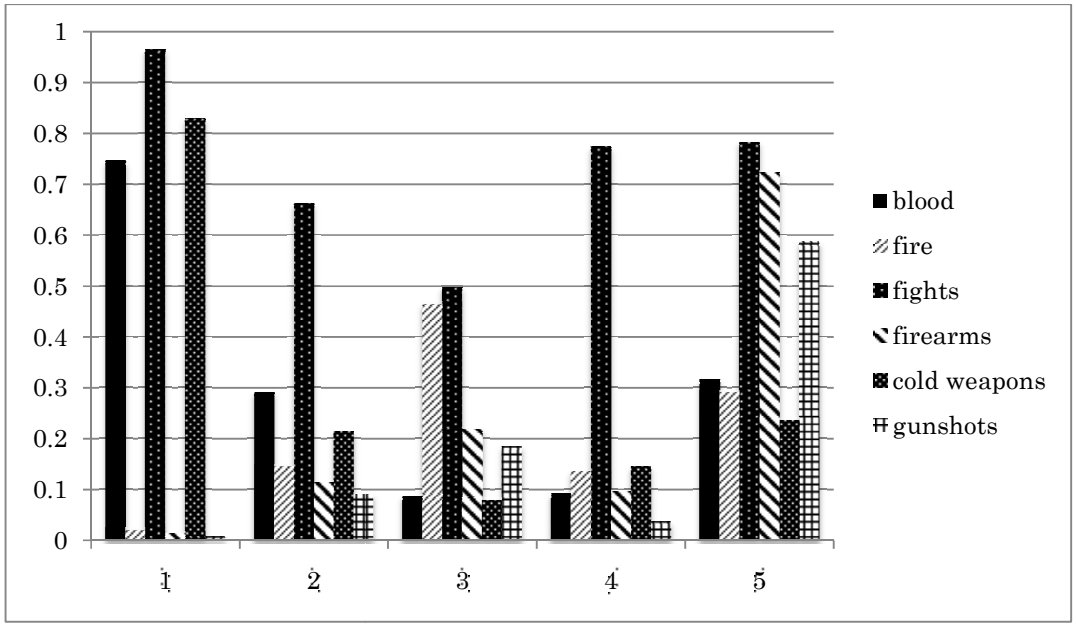

Figure 4. Ratios of segments annotated by each concept for clusters ($K = 50$).

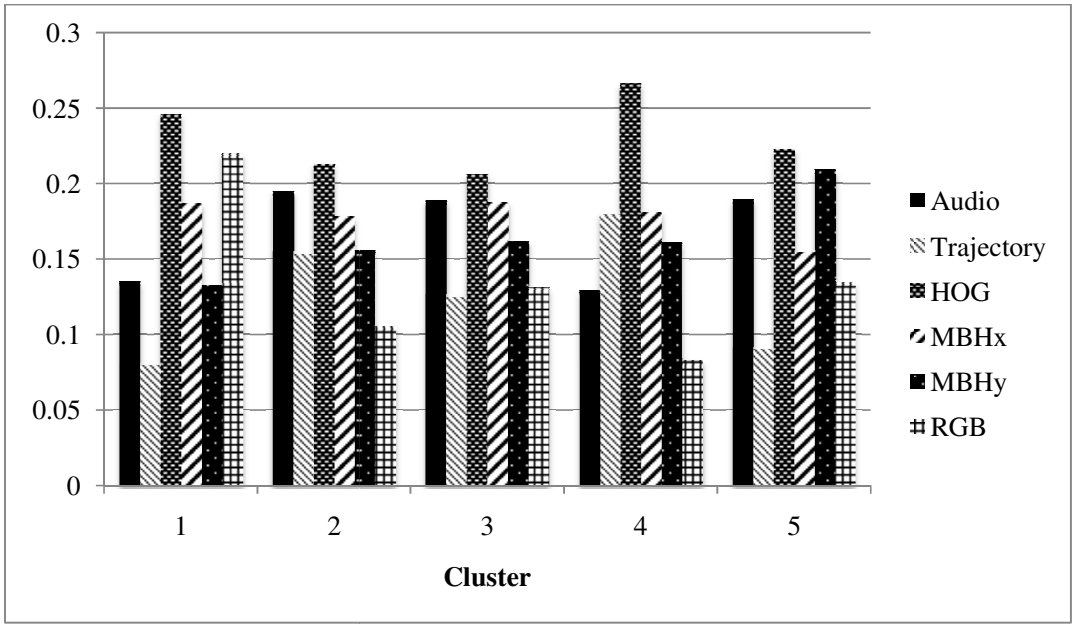

Figure 5. MKL weights for each cluster in Figure 4.

concept by examining the amount of violence-related concepts in it using the cluster number $K = 50$. In MediaEval 2013, participants were also provided with mid-level concepts annotated at frame level by human assessors, which we did not use on our system. They consist of 7 visual concepts: *presence of blood, presence of fire, fights, gory scenes, presence of firearms, presence of cold weapons, carchases,* and 3 audio concepts: *explosions, presence of screams, gunshots*. It should be noted that these mid-level concepts are not always related to violencetags, and often multiple concepts are tagged in one frame. Since they are at frame-level, we converted them to segment-level annotations by simply tagging each segment if half of frames it contains are annotated. The ratio of segments annotated by each mid-level concept in each cluster is shown in Figure 4.Since it is in adequate to display ratios for all clusters and for all mid-level concepts in this figure due to the limit of the available spaces, only 5 representative clusters and 6 concepts are displayed. Additionally, optimized weights trained by MKL for each cluster are displayed in Figure 5.

By studying these figures one can find there are some correlations between concepts and weights. For instance, Cluster 1 has more segments annotated as *blood* and *cold weapons*, and it has higher weights for HOG and RGB feature space at the same time. It is safe to say that this cluster has segments in which fights with *blood* or *cold weapons* appear, just as in *Kill Bill*. Cluster 5, on the other hand, contains a high number of segments tagged as *firearms* and *gunshots,* and its weights for Audio is high. This leads to our presumption that Cluster 5 includes violent scenes of gunfire. On the other hand, the ratio of *fights* is high for all clusters. Though clusters and concepts in MediaEval are unrelated essentially, and so clusters do not always have to be distinctive in Figure 4, this might be caused by the lack of distinctiveness of visual features.

5. CONCLUSIONS

In this paper, we proposed a novel system to detect violent scenes in movies by using Mid-level Violence Clustering and Multiple Kernel Learning with multimodal features. Our experiments proved clustering violence before training is effective, and mid-level concepts can be implicitly learned. Obtained results outperformed the best score in MediaEval 2013 Affect Task. In

addition, we investigated relations between concepts of mid-level clusters and weights trained by Multiple Kernel Learning and found correlations. Future work is to find more appropriate feature vectors, as well as to adopt more suitable clustering method in the context of our system.

ACKNOWLEDGEMENTS

We acknowledge the MediaEval 2013 Affect Task: Violent Scenes Detection http:// www.multimediaeval.org / mediaeval2013/ for providing the dataset that has been supported, in part, by the Quaero Program http://www.quaero.org.

REFERENCES

[1] C. Demarty, C. Penet, M. Schedl, B. Ionescu, V.L. Quang,and Y. Jiang,"The Mediaeval 2013 Affect Task: Violent Scenes Detection," in *MediaEval 2013 Workshop*, Barcelona, Spain, Oct. 18-19 2013.

[2] Liang-Hua Chen, Hsi-Wen Hsu, Li-Yun Wang, and Chih-Wen Su,"Violence Detection in Movies," in *Computer Graphics, Imaging and Visualization (CGIV), 2011 Eighth International Conference*, 2011.

[3] Theodoros Giannakopoulos, Dimitrios Kosmopoulos, Andreas Aristidou, and Sergios Theodoridis,"Violence Content Classification Using Audio Features," in *SETN'06 Proceedings of the 4th Helenic conference on Advances in Artificial Intelligence*, pp. 502–507, 2006.

[4] Fillipe D. M. de Souza, Guillermo C. Chavez, Eduardo A. do Valle Jr., and Arnaldo de A. Araujo,"Violence Detection in Video Using Spatio-Temporal Features," in *SIBGRAPI '10 Proceedings of the 2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images*, pp. 224–230, 2010.

[5] E. Bermejo, O. Deniz, G. Bueno, and R. Sukthankar,"Violence Detection in Video Using Computer Vision Techniques," in *CAIP'11 Proceedings of the 14th international conference on Computer analysis of images and patterns - Volume Part II*, pp. 332–339, 2011.

[6] Jeho Nam, Masoud Alghoniemy, and H. Tewfik, "Audio-Visual Content-Based Violent Scene Characterization," in *International Conference on Image Processing*, Oct. 1998.

[7] Jian Lin and Weiqiang Wang,"Weakly-Supervised Violence Detection in Movies with Audio and Video Based Co-training," in *PCM '09 Proceedings of the 10th Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, pp. 930–935, 2009.

[8] Cedric Penet, Claire-Helene Demarty, Guillaume Gravier, and Patrick Gros, "Multimodal Information Fusion and Temporal Integration for Violence Detection in Movies," in *ICASSP - 37th International Conference on Acoustics, Speech, and Signal Processing*, 2012.

[9] Bogdan Ionescu, Jan Schluter, Ionut Mironica, and Markus Schedl,"A Naive Mid-level Concept-based Fusion Approach to Violence Detection in Hollywood Movies," in *ICASSP - 37th International Conference on Acoustics, Speech, and Signal Processing*, 2012.

[10] Chun Chet Tan and Chong-Wah Ngo,"The Vireo Team at MediaEval 2013: Violent Scenes Detection by Mid-level Concepts Learnt from Youtube," in *Working Notes Proceedings of the MediaEval 2013 Workshop*, Barcelona, Spain, Oct. 18-19 2013, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_12.pdf.

[11] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu,"Action Recognition by Dense Trajectories,"in *IEEE Conference on Computer Vision & Pattern Recognition*, pp. 3169–3176, Colorado Springs, United States, June 2011.

[12] N. Dalal,B. Triggs and C. Shmid, "Human detection using oriented histograms of flow and appearance," in *European conference on computer vision*, 2006.

[13] Stephanie Pancoast and Murat Akbacak,"Bag-of-audio-words Approach for Multimedia Event Classification," in *INTERSPEECH'12*, 2012.

[14] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cedric Bray,"Visual Categorization with Bags of Keypoints," in *Proc. of European Conference on Computer Vision (ECCV)*, pp. 1–22, 2004.

[15] S. Sonnenburg, G. Raetsch, C. Schaefer, and B. Schoelkopf,"Large Scale Multiple Kernel Learning," in *Journal of Machine Learning Research*, 2006.

[16] Francis R. Bach, Gert R. G. Lanckriet, and Michael I. Jordan,"Multiple Kernel Learning, Conic Duality, and the SMO Algorithm," in *ICML '04 Proceedings of the twenty-first international conference on Machine Learning*, 2004.

[17] A. Barla, F. Odone, and A. Verri,"Histogram Intersection Kernel for Image Classification," in *Proceedings of ICIP 2003*, pp. 513–516, 2003.

[18] Soeren Sonnenburg, Gunnar Raetsch, Sebastian Henschel, Christian Widmer, Jonas Behr, Alexander Zien, Fabio de Bona, Alexander Binder, Christian Gehl, and Vojtech Franc, "The SHOGUN Machine Learning Toolbox," in *Journal of Machine Learning Research, 11*, pp. 1799–1802, June 2010.

[19] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze,*Introduction to Information Retireval*,July 2008.

[20] Nadia Derbas, Bahjat Safadi and Georges Quénot,"LIG at MediaEval 2013 Affect Task: Use of a Generic Method and Joint Audio-Visual Words," in *Working Notes Proceedings of the MediaEval 2013 Workshop*, Barcelona, Spain, Oct. 18-19 2013, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_13.pdf.

[21] I. Laptev. "On Space-Time Interest Points,"in *International Journal of Computer Vision*, vol. 64, number 2/3, pp.107-123, 205.

[22] M. Chen and A. Hauptmann,"MoSIFT: recognizing Human Actions in Surveillance Videos,"in *CMU-CS-09-161*, 2009.

[23] H. Liu and P. Singh. "ConceptNet – a practical commonsense reasoning tool-kit,"in *BT Technology Journal*, vol. 22, pp. 211-226, Oct. 2004.

[24] Qi Dai, Jian Tu, Ziqiang Shi, Yu-Gang Jiang and Xiangyang Xue,"Fudan at MediaEval 2013: Violent Scenes Detection Using Motion Features and Part-Level Attributes," in *Working Notes Proceedings of the MediaEval 2013 Workshop*, Barcelona, Spain, Oct. 18-19 2013, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_6.pdf.

[25] Mats Sjöberg, Jan Schlüter, Bogdan Ionescu and Markus Schedl,"FAR at MediaEval 2013 Violent Scenes Detection: Concept-based Violent Scenes Detection in Movies," in *Working Notes Proceedings of the MediaEval 2013 Workshop*, Barcelona, Spain, Oct. 18-19 2013, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_10.pdf.

[26] Ismael Serrano, Oscar Déniz, Gloria Bueno,"VISILAB at MediaEval 2013: Fight Detection," in *Working Notes Proceedings of the MediaEval 2013 Workshop*, Barcelona, Spain, Oct. 18-19 2013, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_11.pdf.

[27] Bruno Do Nascimento Teixeira,"MTM at MediaEval 2013 Violent Scenes Detection: Through Acoustic-visual Transform," in *Working Notes Proceedings of the MediaEval 2013 Workshop*, Barcelona, Spain, Oct. 18-19 2013, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_15.pdf.

[28] Vu Lam, Duy-Dinh Le, Sang Phan, Shin'ichi Satoh, Duc Anh Duong,"NII-UIT at MediaEval 2013 Violent Scenes Detection Affect Task," in *Working Notes Proceedings of the MediaEval 2013 Workshop*, Barcelona, Spain, Oct.18-19 2013, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_27.pdf.

[29] Cédric Penet, Claire-Hélène Demarty, Guillaume Gravier, Patrick Gros,"Technicolor/INRIA Team at the MediaEval 2013 Violent Scenes Detection Task," in *Working Notes Proceedings of the MediaEval 2013 Workshop*, Barcelona, Spain, Oct. 18-19 2013, CEUR-WS.org, ISSN 1613-0073, http://ceur-ws.org/Vol-1043/mediaeval2013_submission_31.pdf.

[30] Technicolor, http://www.technicolor.com, last accessed Nov. 2013.

[31] Hanna M. Wallach,"Conditional Random Fields: An Introduction,"*Technical Report MS-CIS-04-21*, Department of Computer and Information Science, University of Pennsylvania, 2004.

[32] Jianchao Yang, Kai Yu, Yihong Gong and Thomas Huang,"Linear Spatial Pyramid Matching using Sparse Coding for Image Classification,"in *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*,2009.

[33] David Arthur and Sergel Vassilvitskii,"k-means++: The Advantages of Careful Seeding," in *SODA '07 Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027-1035, 2007.

AUTHORS

Shinichi GOTO received his B.Sc in Information and Intelligent Systems from Tohoku University in 2011. He is currently a M.Sc student at Tohoku University. His research interests include video processing, machine learning and web technologies.

Terumasa AOKI is an associate professor at NICHe (New Industry Creation Hatchery Center), Tohoku University. He received his B.E, M.E and Ph.D degree from the University of Tokyo, in 1993, 1995 and 1998 respectively. He has received various academic excellent awards such as Young Scientist Award from MEXT (the Ministry of Education, Culture, Sports, Science and Technology in Japan, in 2007), Yamashita Award and the Best Education Award from IPSJ (Information Processing Society of Japan, in 2001 and 2007 respectively), two Best Paper Awards from IIEEJ (the Institute of Image Electronics Engineers of Japan, in 2004 and 2009) etc. He is well known as the developer of DMD (Digital Movie Director) and NeoPoster etc. His current research topic is digital content technology, especially image processing, computer vision and computer graphics.