Affective and Holistic Approach at TRECVID 2010 Task - Semantic Indexing (SIN)

Kok-Meng Ong, Supheakmungkol Sarin and Wataru Kameyama
Graduate School of Global Information and Telecommunication Studies
Waseda University
1011 Nishi-Tomida, Honjo, Saitama 367-0035 Japan
Email: ongkokmeng@aoni.waseda.jp, mungkol@fuji.waseda.jp, wataru@waseda.jp

***Abstract*—This paper reports our experiments for TRECVID 2010 task: Semantic Indexing. We present two approaches namely, Affective and Holistic. In the first approach, we have used combination of affective features from image, video and audio trained with neural network algorithm. Image features employed are color histogram and face detection from the keyframe. The number of face is also used in one of the runs. Video features include the motion activity and shot duration. Additionally, the audio power is included as feature. For the second approach, color, texture and scene features are extracted from the whole keyframe image as well as its background and saliency regions. Genetic algorithm is used to find the weight of each feature for effective combination. Then, KNN is used to propagate the annotation. We have submitted 4 runs where we distinguish the first two as affective category and the the last two as holistic ones. The summary is as follows:**

- ***kmlabGITS1*-color histogram, motion, rhythm, sound and face number trained using neural network**
- ***kmlabGITS2*-color histogram, motion, rhythm, sound and without face number trained using neural network**
- ***kmlabGITS3*-combination of 5 image features (hsv_bg, $gabor$, $haar$, $gist$ and lab_bg) using Genetic Algorithm and KNN**
- ***kmlabGITS4*-combination of 5 image features (hsv, hsv_bg, $haar$, $haar\_roi$ and $gist$) using Genetic Algorithm and KNN**

I. Introduction

This paper reports our experiments for TRECVID [1] 2010 task: Semantic Indexing. We have participated in the Lite Version of the task, where the purpose is to detect the following concepts in the test dataset [2]:

- [004] Airplane_Flying
- [015] Boat_Ship
- [019] Bus
- [028] Cityscape
- [029] Classroom
- [041] Demonstration_Or_Protest
- [059] Hand
- [084] Nighttime
- [105] Singing
- [117] Telephone

This paper is organized as follows. In the next section, we describe the features that we have used in the runs, that includes the image, video and sound features. Section III outlines our training algorithm. Our approaches for the 4 runs that were submitted for TRECVID evaluation are described in Section IV. The evaluation result is presented in Section V. We present our discussion in Section VI and finally the paper is concluded in Section VII

II. Feature Extraction

A. Image Features

The keyframe for each shot, provided by TRECVID, is used to extract image features that represent the shot. Utilizing the image, the following features are extracted.

1) Affective image features:

- Color Histogram: The basic color histogram is extracted from the keyframe and used as one of the input feature. The keyframe is first converted into the HSV planes, and histogram of 5 bins for each of the planes is calculated. The histogram is then normalized and used as the input features.
- Face Detection The number of faces is extracted as input feature. The face detection technique based on Viola-Jones detector [3] is used. The pre-trained objects used for the Haar detector that is provided by OpenCV [4] is used. The total number of face that is detected in the keyframe using this detector is used as the input feature.

2) Holistic image features: Human exhibits the exquisite ability at rapidly identifying the gist of the scene of the image. Usually, a human observer of an image at a fraction of second can summarize the essential information about the image such as indoor/outdoor, street, beach, landscape, etc. [5], [6]. Saliency is also a very important point of interest when human observes image because they tend to focus on some important regions or ROIs. Study has shown that the concurrent use of gist of the scene and saliency is a major trait of human vision system [7]. These give reasons for our idea.

In this experiment, we would like to capture these important features in addition to the basic ones (color and texture from the whole image) as proposed in [8]. The original research on gist of the scene has been reported in [9] with quite a successful rate. For saliency detection, Itti et al.'s work [10] has been the most popular one. However, it is rather complex and computationally expensive. Two recent approaches introduced by Hou et al. and Achanta et al. in [11], [12] are simple and yet give good performance in real-time computation. Fig. 1 shows the overall view of feature extractions.

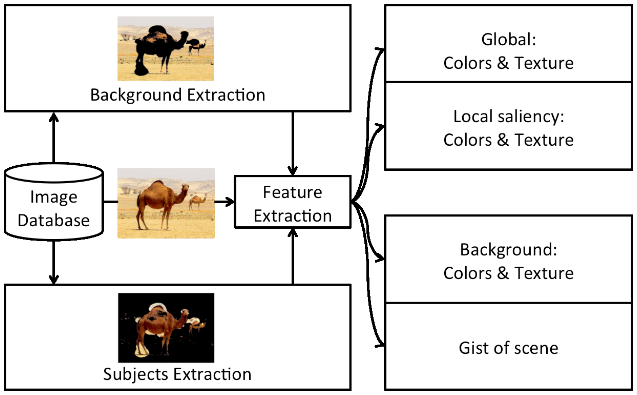

Fig. 1. Holistic Feature Extraction Process

- Color and Texture
  First, the image is processed to extract the background and subjects. To do this, we leverage the combination of the recent approaches of saliency extraction using spectral residual model and frequency-tuned model as seen in [11], [12]. In our implementation, the threshold for saliency region of each model is computed as follows.

  $$Cutsize(I) = mean(SMap(I)) + std(SMap(I)) \quad (1)$$

  where $SMap(I)$ is the saliency map of image I. The final subject and background areas are the union and the intersection of the area calculated from each model respectively. Next, a number of color and texture features are extracted from the original as well as background and subject areas. We compute the color histogram of the saliency regions for the three color spaces namely, RGB, LAB and HSV and two wavelet textures namely, Haar and Gabor.
  - RGB, LAB, HSV: are simple color histogram in the respective color spaces and computed in 3 channels each with 16 bins.
  - Gabor: a three scales and four orientations filter is used. Then, each response images are split into non-overlapping rectangular blocks. We calculate the mean filter response magnitudes from each block over all twelve response images.
  - Haar: a two by two edge filter is used. The wavelet responses are generated by block-convolution of an image with Haar filters at three different orientations (vertical, horizontal and diagonal). Convolution with a sub-sampled image are conducted at different scales. Afterward, the image is rescaled to size 64x64 pixels, a Haar feature is generated by concatenating the Haar response magnitudes.
- Gist of scene: The gist descriptors describe the spatial layout of an image using global features derived from the spatial envelope. It is shown to be very good in scene categorization, we use the original implementation in [9] and compute the descriptors at 256x256.
- Feature Normalization: Given that the feature is extracted, normalization is needed. For this, we chose *Rank Normalization* and uniformly scale feature values to $[0, 1]$ range as follows. Let x_1, x_2, ..., x_n be sample for a feature component of all images, first we find the order statistics $x_{(1)}$, $x_{(2)}$, ..., $x_{(n)}$ and then we replace each image's feature value by its corresponding normalized rank as

  $$\tilde{x_i} = \frac{\underset{x_1, x_2, \ldots, x_n}{Rank}(x_i) - 1}{n - 1} \quad (2)$$

  where x_i is the feature value for the i^{th} image.

B. Video Features

Two video features are extracted from the video. The video features are the extracted based on the arousal model proposed by Hanjalic et. al. [13].

1) Motion: Motion activity is actively researched in the field of affective video processing. The viewers's emotion is said to be influence by the amount of motion activity in video [13]. Therefore, we have extracted motion activity as one of the features for this task.

In order to extract the motion feature for each shot, the motion activity $m(k)$ for each frame is calculated. First, the Motion vector for macroblock (16×16 pixels), $\vec{v_i}(k)$, between frame k and $k + 1$ is calculated. Then the motion activity is obtained by the sum of all normalized motion vector as in the following equation:

$$m(k) = \frac{1}{B|\vec{v}_{max}|}\left(\sum_{i=1}^{B} |\vec{v}_i(k)|\right) \quad (3)$$

where B is the total number of motion vectors of the frame, and $|\vec{v}_{max}|$ is the value of maximum motion vector within frame k.

Two criteria, Compatibility and Smoothness has to be satisfy as the features that represent the arousal model [13]. This is because each extracted video features is unique and have different scales. Therefore, for the purpose of comparison between the features, the Compatibility criterion has to be satisfied. Furthermore, due to several characteristic listed below, the Smoothness criterion has to be satisfied too:

- The rapid change of the value of equation 3, even in the same shot.
- The difference of the motion activity for consecutive shots.
- The noise that is unavoidable when the motion activity is calculated on frame basis. Therefore, smoothing the curve has the effect of removing noise.

In order to satisfy the above-mentioned criteria, equation 4 is used as follow:

$$Motion(k) = \frac{\max(m(k))}{\max(\widetilde{m}(k))}\widetilde{m}(k) \quad (4)$$

where $\widetilde{m}(k) = m(k) * K(l_1, \beta_1)$ is the convolution between $m(k)$ and Kaiser window $K(l_1, \beta_1)$. l_1 and β_1 is set at 700 and 5 [13].

2) Rhythm: Another video feature that is incorporated in our run is the rhythm feature [13]. According to [13], one of the way the video makers use to change the tempo or rhythm of video is by changing the length of the video shot. For example, in order to increase the tempo of some action scene, normally the shot length is shortened, while to reduce the pace of the video for some narration, the shot length is lengthen. Therefore, the shot length is used in the following equation to obtain the rhythm feature of the video.

$$c(k) = e^{((1-(n(k)-p(k)))/\delta)} \tag{5}$$

where $p(k)$ and $n(k)$ are the frame index for the first frame of the current and next shot correspondingly, and δ is a constant that is used to alter the overall value of $c(k)$. Here, δ is set as 300 [13]. In addition, because $c(k)$ is a step function, in order to satisfy the Compatibility and Smoothness criteria mentioned before, equation 6 is applied:

$$Rhythm(k) = \frac{\max(c(k))}{\max(\widetilde{c}(k))} \widetilde{c}(k) \tag{6}$$

Similar to equation 4, $\widetilde{c}(k) = c(k) * K(l_1, \beta_1)$ is the convolution of $c(k)$ and Kaiser window $K(l_1, \beta_1)$. l_1 and β_1 are set to 700 and 5 [13].

The average value across the shots for both Motion and Rhythm is used as shot features for the runs.

C. Audio

One audio feature is the extracted based on the arousal model proposed by Hanjalic et. al. [13] as follow:

1) Sound: The audio effect in the video has influence towards viewer's affective response [13]. The loudness of sound (high audio energy), and high speech tempo influence the arousal level of viewers while the articulation, and music often influence the valence of viewers. Here, the audio energy is extracted as the sound feature in the runs. In order to extract the sound energy for each video frame, the audio sample in a frame s is taken as the ratio of audio sampling rate and the frame rate. Then for sound energy $e(k)$ of frame k, the power spectral is taken and its sum is taken as the sound energy value for the frame.

Here, the Kaiser window is employed again to obtained $\widetilde{e}(k) = K(l_1, \beta_1) * e(k)$. l_1 and β_1 is set to 700 and 5 [13]. However, the sound feature can not be determined with only regards to the energy level. For example, a shot with low average energy, but with a few energy peak is deemed to be more arousal if compare and monotonous shot with averagely high sound energy. Therefore, in order to obtain sound feature, a weight is added as in equation 7 below:

$$Sound(k) = e_n(k)(1 - \overline{e}_n) \tag{7}$$

In other words, the smoothen sound energy $\widetilde{e}(k)$ is normalized $e_n(k) = \frac{\widetilde{e}(k)}{\max(\widetilde{e}(k))}$, then weight $(1 - \overline{e}_n)$ is applied to $e_n(k)$ in order to obtain the final sound feature. Where $\overline{e}_n = \frac{1}{W} \sum_k e_n(k)$, and W is the length of the video.

Again, the average value across the shots for Sound is used as shot features for the runs.

III. Training Algorithm

The training dataset we have used is the IACC training dataset [2]. Algorithms below are separated for each category of runs.

A. Affective runs

1) Neural Network: Neural Network is used as classifier. Multilevel perceptron neural network with the input nodes corresponding to the number of input features, and output node for each concept to be detected is used. The network is trained for each concept based on the IACC training dataset. The output of the trained network on each test data is then used as the measure to rank the shot to each concept.

B. Holistic runs

1) Concept propagation: We calcuate the distance between images. The L1 or block distance is used for all the features. We use the K Nearest Neighour (KNN) model to propagate the concepts. The first concepts are selected from the nearest neighbor. If more concepts are needed, they are selected from neighbors 2 through N based on co-occurrence and frequency.

2) Feature selection: We employ a simple expand/reduce algorithm in order to first find the best set of features among all the features extracted. For this process, each feature contributes equally towards the image distance. Let $d(i, j)$ be combined distance of image Ii and Ij. If $\tilde{d}^k_{(i,j)}$ is the scaled distance, then

$$d(i, j) = \frac{1}{N} \sum_{K=1}^{N} \tilde{d}^k_{(i,j)} \tag{8}$$

We arrive at the following two best sets of combination that we use for run 3 and 4. They are ranked in order.

– hsv_bg, $gabor$, $haar$, $gist$, and lab_bg
– hsv, hsv_bg, $haar$, $haar\_roi$, and $gist$

Note: *feature_bg* and *feature_roi* are features extracted from the background and regions of interest (or saliency regions) respectively.

3) Genetic Algorithm: Ultimately, we would like to find the optimal combination of these best features. Therefore, we need to know the weighting of each features in the following equation:

$$d(i, j) = \frac{1}{N} \sum_{K=1}^{N} w_k * \tilde{d}^k_{(i,j)} \tag{9}$$

where w_k is the weight of the feature k.

To do this, we run a genetic algoritm with the following setting:

– Mutation: 10%
– Elite: 20%

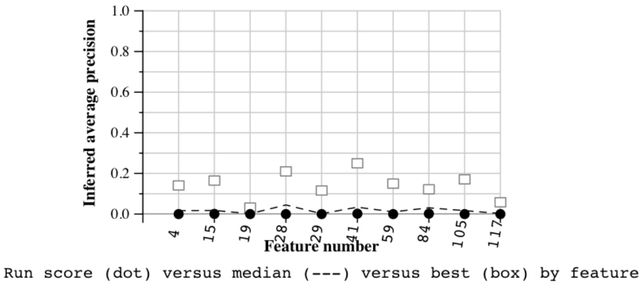

Fig. 2. Result for Run 1 : kmlabGITS1

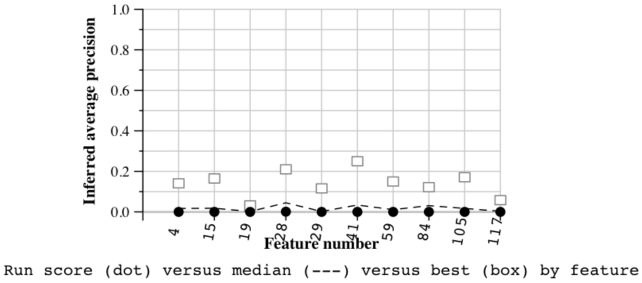

Fig. 3. Result for Run 2 : kmlabGITS2

- Cross over: 20%
- Number of population: 20
- Number of generation: 5

IV. Our Approach

A. Run 1 : kmlabGITS1

For this run, the input features that are included: color histogram, motion, rhythm, sound and face number as explained in the earlier section. A neural network classifier is trained using the IACC training data. The test shot's input features are fed into the trained network and the output is used to rank the shot for each concept.

B. Run 2 : kmlabGITS2

For this run, the input features that are included: color histogram, motion, rhythm, sound. Similarly, neural network is employed in this run. The only difference with Run 1 is: Run 2 is trained without the face number as input feature.

C. Run 3 : kmlabGITS3

A combination of 5 image features are used in this run (hsv_bg, $gabor$, $haar$, $gist$, and lab_bg). As presented earlier, our approach is based on annotation propagation. We first calculate the average number of concepts for one keyframe image based on the training set. Then, we propagate the concepts from the training set to the test set using the KNN method. Features are combined using the weights output from the genetic algorithm. Finally, Singular Vector Decomposition (SVD) technique is used to select the top 2000 keyframes for each concepts.

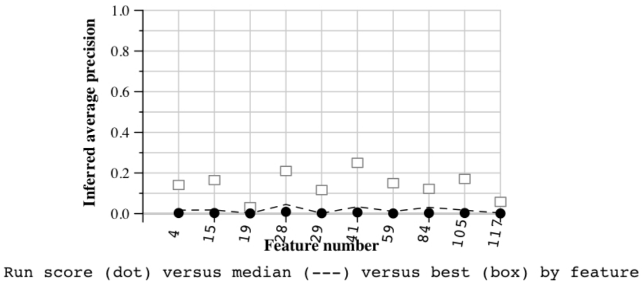

Fig. 4. Result for Run 3 : kmlabGITS3

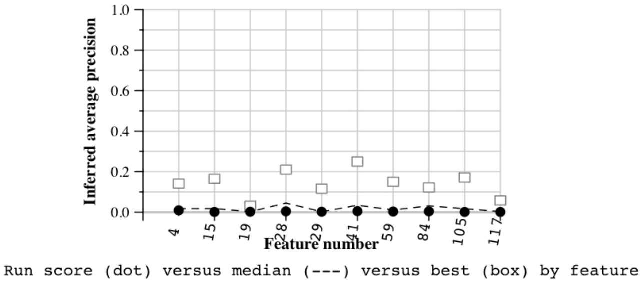

Fig. 5. Result for Run 4 : kmlabGITS4

D. Run 4 : kmlabGITS4

We use the same method as the previous run but with a combination of 5 other image features (hsv, hsv_bg, $haar$, $haar\_roi$, and $gist$).

V. Result

Based on the evaluation by TRECVID, the inferred total true shots for the 10 concepts in Light version is 14987. Out of these inferred true shots, the performance of our run are as below:

A. Run 1 : kmlabGITS1

Inferred true shots returned by this run is 266. With the inferred precision at depth 10 shots as 0.030, 100 shots at 0.024, 1000 shots at 0.018 and 2000 shots at 0.013.

The breakdown of results for the 10 concepts in Light version is shown in Fig. 2.

B. Run 2 : kmlabGITS2

Inferred true shots returned by this run is 193. With the inferred precision at depth 10 shots as 0.010, 100 shots at 0.013, 1000 shots at 0.008 and 2000 shots at 0.010.

The breakdown of results for the 10 concepts in Light version is shown in Fig. 3.

C. Run 3 : kmlabGITS3

Inferred true shots returned by this run is 902. With the inferred precision at depth 10 shots as 0.050, 100 shots at 0.037, 1000 shots at 0.041 and 2000 shots at 0.045.

The breakdown of results for the 10 concepts in Lite version is shown in Fig. 4.

D. Run 4 : kmlabGITS4

Inferred true shots returned by this run is 881. With the inferred precision at depth 10 shots as 0.020, 100 shots at 0.052, 1000 shots at 0.032 and 2000 shots at 0.044.

The breakdown of results for the 10 concepts in Lite version is shown in Fig. 5.

VI. Discussion

Our submission of 4 runs can be broken down into two approaches: the first approach includes Run 1 and Run 2, while second approach includes Run 3 and Run 4.

For the first approach, Run 2: *kmlabGITS2* is the baseline run which includes basic color information and features based on arousal model [13], as input features. Neural networks are trained for each concepts based on the training data. Run 1: *kmlabGITS1* added the number of faces appears in the keyframe as added input feature. The inclusion of face number in Run 1 increases the number of inferred true shots. The inclusion of face number as feature possibly improves the performance on concepts with appearance of human like [041] Demonstration_Or_Protest and [105] Singing. However, from the overall point of view, the performance of the first approach (Run 1 and Run 2) is not satisfactory if compared to the second approach (Run 3 and Run 4). Our first approach focuses mainly on the affective aspect of the features. This approach might not be appropriate in detecting concrete concepts in the task, which lead to the poor performance.

For the second approach, though the results are not satisfactory. We believe that our method can be used for this task. There are obviously rooms for improvement. First, the expand/reduce algorithm that we utilize is very simple and not optimal. That might be the reason that we only obtain a limited number of features in the two sets. Second, due to large training set and time constraint, the genetic algorithm was scaled down to a limited environment setting (very small number of population and number of generation). Third, also because of the previous reason, we did not use the full training dataset, only the training dataset that has the 10 light concepts were used. Moreover, we did not exploit the negative concepts that associated with the trainning dataset at all. We believe that considering all these aspects, we can achieve better results in the future.

VII. Conclusion

We report our experiments combining different features of image, video and audio with different models for the task of semantic indexing. Though, we could not achieve satisfactory results, these experiments give us insight on our features and methods employed. One of the reasons is that some affective features are not suitable for the concepts that we would like to detect. Additionally, we did not fully exploit the training dataset. We also would like to explore more on the combination of these features especially the cross-combination between affective and holistic features. Moreover, we would like to further investigate on other advanced features and classifiers that could be incorporated to this task. These define our future works.

References

[1] A. F. Smeaton, P. Over, and W. Kraaij, “Evaluation campaigns and trecvid,” in *MIR ’06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*. New York, NY, USA: ACM Press, 2006, pp. 321–330.

[2] “Trecvid 2010 training data,” Website, 2010, http://www-nlpir.nist.gov/projects/tv2010/tv2010.html#data.

[3] P. Viola and M. Jones, “Rapid object detection using a bossed cascade of simple featuress,” in *Computer Vision and Pattern Recognition 2001, CVPR 2001*, 2001, pp. I–511–I–518.

[4] G. Bradski and A. Kaehler, *Learning OpenCV Computer Vision with the OpenCV Library*, 1st ed. O’Reilly Press, 2008.

[5] M. C. Potter, “Short-term conceptual memory for pictures.” *Journal of experimental psychology Human learning and memory*, vol. 2, no. 5, pp. 509–522, 1976. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/1003124

[6] A. Friedman, “Framing pictures: the role of knowledge in automatized encoding and memory for gist.” *Journal of experimental psychology General*, vol. 108, no. 3, pp. 316–355, 1979. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/528908

[7] C. Siagian and L. Itti, “Biologically inspired mobile-robot self localization,” *The Neuromorphic Engineer*, pp. 1–2, Dec. 2007.

[8] A. Makadia, V. Pavlovic, and S. Kumar, “A New Baseline for Image Annotation,” in *ECCV (3)*, 2008, pp. 316–329.

[9] A. Oliva and A. Torralba, “Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope,” *International Journal of Computer Vision*, vol. 42, pp. 145–175, 2001.

[10] L. Itti, C. Koch, and E. Niebur, “A Model of Saliency-Based Visual Attention for Rapid Scene Analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, pp. 1254–1259, 1998.

[11] X. Hou and L. Zhang, “Saliency Detection: A Spectral Residual Approach,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition CVPR ’07*, 2007, pp. 1–8.

[12] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned salient region detection,” *IEEE Conference on Computer Vision and Pattern Recognition (2009)*, vol. pages, no. Ic, pp. 1597–1604, 2009. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5206596

[13] A. Hanjalic and L. Q. Xu, “Affective video content representation and modelling,” *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 143–154, Feb 2005.