Abstract

Recently more and more researches focus on the concept extraction from unstructured video data. To bridge the semantic gap between the low-level features and the high-level video concepts, a mid-level understanding of the video contents, i.e., salient object is detected based on the techniques of image segmentation and machine learning. Specifically, 21 salient object detectors are developed and tested on TRECVID 2005 development video corpus. In addition, a boosting method is proposed to select the most representative features to achieve a higher performance than only using single modality, and lower complexity than taking all features into account.

1. Introduction

With the prevalence of the digital cameras and the development of communication techniques, more and more videos are becoming available. Thus video retrieval over large-scale collections is becoming more and more important. However, query-by-example (QBE) of video is inconvenient for most of users because the representative query examples may usually be unavailable. In addition, it is very difficult to automatically find the users’ interestingness. Thus querying by semantics, which can specify the query concepts via keywords, can probably better meet the user’s need.

However, the extraction of the semantics of the video data is also a hard work for the researchers because of the varieties of video concepts. Focusing on some specific video domains may be able to avoid this problem [1] [17]. For example, soccer videos are analyzed and features with a lot of human prior knowledge, such as the playfield region, the camera motion, and the players’ position, are extracted to describe the characteristic of soccer videos in [1]. In [17], pairwise constraints are added into the regularized empirical risk to deal with problem of insufficient labeled data in video object classification, which is tested on two surveillance video data sets.

A promising way to narrow the semantic gap is to build large-scale concept ontology for the video [12]. In [16], five graphical models are used for learning the multi-concept relation to improve the performance of the concept detection. Four different mining models are proposed in [15] to reveal semantics beyond the discovered patterns. And an event, which is regarded as stochastic process in the semantic concept space [6], is a specific pattern in which several concurrent concepts evolve.

So it is important to develop the basic concept detectors to build the hierarchy of the ontology of the video. Naphade et al. [9][10] build some models of regional semantic concepts with four features, including color correlogram, edge orientation histogram, co-occurrence texture, and moment invariants, at four different granularities. Regions having the similar semantics can group together by Self-Organization Map (SOM) learning in [20]. And in [18] [19], the framework of semi-supervised learning is used to detect objects in aerial imagery. Multiple-instance learning is also taken by some researchers [2] [11]. In [11], a negative hypothesis is built from the instances in negative bags. And then the positive hypothesis is built with the instances in positive bags with weights evaluated by the negative hypothesis. Chen et al. [2] embeds bags into an instance-based feature space and selects the most important features to identify instances that are relevant to the observed classification.

In this paper, we develop 21 salient object detectors which are defined by the Large Scale Concept Ontology for Multimedia (LSCOM). We label the regions which are segmented from the key frames of the video shots of TRECVID 2005 development data set. Eleven regional features, including color, texture, edge, and shape, are extracted. And supporting vector machines are trained on each feature. The final results are obtained by the proposed boosting method which adjusts the weight of each classifier iteratively in a framework similar to AdaBoost.

The rest of paper is organized as follows: We give the definition of the salient object in Section 2 and introduce the detection function in Section 3. Experiments and their analysis are shown in Section 4. In Section 5, we give a conclusion and future work plan.

2. Salient Object

Semantics are expected to be extracted from the videos for better indices of the videos. And it’s also helpful for
retrieving the information we need which is distributed on different kinds of modalities of data, such as visual, auditory, and textual. For example, a query of “George Bush” may include his appearance on TV, his speech on radio, and his news on newspaper or web. However, the semantics of a video clip are based on the complicated architecture of human knowledge, whereas the low-level features are based on the signal processing which we can easily get from the video directly. The low-level features of different video clips with the same semantic may vary in a wide range. There is a huge difference between them which is called “semantic gap”. To bridge the semantic gap, a middle-level representation, which is called “salient object”, is first proposed to express the video content efficiently by Fan et al. [7][8].

In [8], the salient objects are defined as the visually distinguishable video components that are related to human semantics. According to this definition, we define 21 salient objects to detect based on TRECVID 2005 development video data, which contains more than 70,000 shots. They are listed in the following table:

<table>
<thead>
<tr>
<th>Airplane</th>
<th>Animal</th>
<th>Boat_Ship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bus</td>
<td>Building</td>
<td>Computer_TV-screen</td>
</tr>
<tr>
<td>Charts</td>
<td>Car</td>
<td>Explosion_Fire</td>
</tr>
<tr>
<td>Desert</td>
<td>Flag-US</td>
<td>Flowers</td>
</tr>
<tr>
<td>Hand</td>
<td>Maps</td>
<td>Mountain</td>
</tr>
<tr>
<td>Road</td>
<td>Sky</td>
<td>Snow</td>
</tr>
<tr>
<td>Truck</td>
<td>Vegetation</td>
<td>Waterscape_Waterfront</td>
</tr>
</tbody>
</table>

Table 1. Salient objects to detect.

The labeling results provided by NIST and LSCOM are simply judgments on whether a key frame of a shot contains a concept or not. In order to describe the location more accurately, image segmentation is taken on all key frames. The parameters of the segmentation function are adjusted so the images are a little over-segmented, and one region contains as fewer salient objects as possible. The ideal segmentation is that one region contains only one salient object. Yet in practice, if most part of a region is a kind of salient object, the region is labeled as positive for that object. The manually labeling work is based on the official annotations, which means that given a salient object, the assessors will only label the regions of a small portion of shots which contain that salient object.

Besides the visual salient objects, we consider the definition of salient objects can extend to the auditory field, although it’s really hard to label them for the usual overlapping and different starting and ending time of these auditory salient objects. But without doubt, it can be of much help in the description of a shot. And we leave it to our future work.

3. Detection Function

3.1. Image Segmentation

There are many segmentation methods based on different criteria [3][5][13]. Specially, NCut [13] transforms the image segmentation problem to an eigen problem, while JSeg [5] defines a J-Value, which considers both the color and texture information of the image, and looks for a segmentation solution to minimize the J-Value of the segmented image. And mean shift applied in image segmentation [3], is a procedure of discontinuity preserving smoothing and grouping. In this paper, JSeg is adopted for its good performance and low complexity. Some segmentation results are shown in Figure 1. Over-segmentation is achieved for the purpose mentioned in Section 2.

3.2. Feature Extraction

Given a region of an image, 11 features are extracted to characterize the visual properties of this region. They are: (a) the average color and the color variance in L*A*B* color space with 7 dimensions (the average and the variance on each channel and the variance using Euclidean distance on three channels); (b) Tamura texture with 15 dimensions (the coarseness, the contrast and some other histograms); (c) four 5-dimension features of the co-occurrence matrix on the horizontal, vertical, 45 degree, and 135 degree directions, respectively; (d) the color layout with 12 dimensions, defined in MPEG-7; (e) the scalable color with 64 dimensions, defined in MPEG-7; (f) the edge histogram with 5 dimensions, similar with the edge histogram descriptor defined in MPEG-7; (g) 4-dimension information of the circumscribing rectangle and the coverage ratio of the region; (h) 7-dimension invariant moments.

3.3. Classifier Training

Supporting vector machine (SVM) has a good performance on 2-class learning problem. For a given salient object, if most of a region \( x_i \) is, or is part of, this salient
if most of a region $X_i$ is, or is part of, this salient object, its label $y_i = +1$. There are some transformation parameters $W$ and $b$ such that $f(X_i) = W \cdot \Phi(X_i) + b \geq +1$. And for the negative region, we have $f(X_i) = W \cdot \Phi(X_i) + b \leq -1$. $\Phi(X_i)$ is called kernel function which maps $X_i$ into higher dimensional space. We adopt the radial basis function (RBF) as the kernel function in our experiment due to the effective approximation of the geometric property of each homogeneous regional feature of the image, which is $\mathcal{R}(X_i, X_j) = \exp(-\gamma |X_i - X_j|)$.

The SVM training is actually the following optimization problem:

$$
\min \left\{ \frac{1}{2} W^T W + C \sum_{i=1}^{m} \xi_i \right\}
$$

where $\frac{1}{2} W^T W$ stands for the reciprocal of the margin between two supporting planes and $\xi_i$ stands for the training error rate. $C > 0$ is the penalty parameter to adjust the margin of supporting planes and the training error rate. The parameter pair $(C, \gamma)$ affects the performance of the model a lot. N-folder cross-validation and grid-search on a wide range are taken to obtain the optimal one.

### 3.4 Classifier Combination

Each homogeneous feature characterizes a certain kind of visual properties for different regions of the images, and different salient objects may have different representative features, which behaves as the classifiers trained from different features have different accuracies. It is thus very important to take some techniques to tune the weight distribution of the classifiers, which is actually a procedure of feature selection, for different types of salient object to achieve a higher performance.

Linear regression is a regress method of using linear function to combine some variables for an expected value. In our framework, the expected value is the label of a region and the variables are the confidences on different features. That is:

$$
C(R_i) = \sum_{j=1}^{m} w_j C_j(R_i) + b
$$

where $m$ is the number of features, $w_j$ is weight of the $j$-th feature, $C_j(R_i)$ is the confidence of $R_i$ on the $j$-th feature and $b$ is the bias. In fact, linear regression is the same with one-layer neuron networks, and $w_j$ and $b$ can be obtained by an algorithm named Adaline [14], which modifies $w_j$ and $b$ iteratively until convergence. It is proved to converge to the least squares solution.

The proposed boosting algorithm, which is similar to AdaBoost, is also a method, which can select the most representative features for the detection of a given salient object. The detail of the algorithm is described as follows:

1) Assume that there are totally $m$ classifiers trained on $m$ features respectively. For a region $r_i = ((x_{i1}, x_{i2}, \ldots, x_{im}), y_i)$, $y_i$ is its label and $(x_{i1}, x_{i2}, \ldots, x_{im})$ are the confidences of $r_i$ to be positive, output by the classifier trained on different features, each dimension of which ranges from 0 to 1. Training set $R = \{r_1, r_2, \ldots, r_n\}$ with a distribution $D = \{d_1, d_2, \ldots, d_n\}$, $d_i = \frac{1}{n}$. And initial weights of each classifier, $w_{1,1} = w_{1,2} = \ldots = w_{1,m} = \frac{1}{m}$. Let $t = 1$.

2) Measure the goodness of the $j$-th feature at the $t$-th times,

$$
g_{j,t} = \begin{cases} 
\frac{\sum_{i=1}^{n} d_i x_{ij}}{D_j}, & \text{if } y_i = +1 \\
\frac{\sum_{i=1}^{n} d_i (1 - x_{ij})}{D_j}, & \text{if } y_i = -1 
\end{cases}
$$

and the goodness of the ensemble at the $t$-th times,

$$
G_t = \sum_{j=1}^{m} w_{t,j} g_{j,t}.
$$

3) Update

$$
d_i = \begin{cases} 
\frac{d_i \cdot \exp(0.5 - \sum_{j=1}^{m} w_{t,j} x_{ij})}{Z_D}, & \text{if } y_i = +1 \\
\frac{d_i \cdot \exp(\sum_{j=1}^{m} w_{t,j} x_{ij} - 0.5)}{Z_D}, & \text{if } y_i = -1 
\end{cases}
$$

where $Z_D$ is the normalization factor.

4) Update $w_{t+1,j} = \frac{w_{t,j} \cdot e^{g_{j,t}}}{Z_g}$, $Z_g$ is the normalization factor.

5) $t = t + 1$, repeat from the 2nd step until $t = T$.

6) The final classifier $C(r) = \sum_{i=1}^{T} \sum_{j=1}^{m} G_{t,i,j} C_j(r)$

where $C_j(r)$ is the confidence of region $r$ on $j$-th feature and $Z$ is also the normalization factor.

### 3.5 Object Detection

After shot boundary detection, key frames of the shot are segmented into regions, and eleven regional features referred before are extracted and input into the SVM models respectively after normalization. The final confidence of a region for a given salient object is the boosting result of classifiers trained on each feature. The region is recognized as the salient object with the highest confidence which is also above an empirical threshold. If not, this region can not
be recognized. As shown in Figure 2, the detection is satisfactory.

![Figure 2. Salient object detection of video shots.](image)

4. EXPERIMENTS

Our experiments are conducted on TRECVID 2005 development video data, which contains more than 70,000 shots. And 2/3 of these shots are used for training, 1/3 for testing. To evaluate the performance more conveniently, an index considering both precision and recall is taken, which is defined as:

\[
F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

4.1. Comparison of Combination Methods

Three different combination methods are compared: each classifier has the same weight, linear regression and our boosting algorithm. The comparison is shown in Figure 3. Linear regression gets an obvious worse performance because its target is the least squares error thus not suitable for this unbalanced dataset in which there are much more negative samples than the positive. From Table 2, we can find out that detectors with the proposed boosting method have got the highest maximum, minimum and average \(F_1\) performance and the lowest variance which means the stability of their performances are best.

![Figure 3. Experiment results for different combination methods.](image)

<table>
<thead>
<tr>
<th>Combination Methods</th>
<th>Max</th>
<th>Min</th>
<th>Avg</th>
<th>Var</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same Weight</td>
<td>0.86</td>
<td>0.49</td>
<td>0.73</td>
<td>0.008</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>0.81</td>
<td>0.36</td>
<td>0.60</td>
<td>0.018</td>
</tr>
<tr>
<td>Boosting</td>
<td>0.88</td>
<td>0.54</td>
<td>0.74</td>
<td>0.006</td>
</tr>
</tbody>
</table>

Table 2. Statistical comparison of combination methods.

4.2. Boosting Iteration Times

In this experiment, the effect of the iteration times \(T\) in the boosting method is concerned. From Figure 4, we can find out that only 4 salient objects benefit a little from the increase of the iteration times while others keep stable, which shows that the speed of convergence of boosting is satisfactory.

4.3. Number of Features

After boosting, some features with the smaller weights are discarded to verify the effectiveness of the increase of features. Figure 5 shows that most of the features in the feature set are not too useful for the performances of the most of salient object detectors, because increasing or discarding features doesn’t affect obviously, except for when there is only one feature left. We can see that the performances dramatically decrease when 10 features are
ances dramatically decrease when 10 features are discarded, while only one single feature is left. This feature has the largest weight which means it is the best among the eleven features. It demonstrates that the multiple-feature fusion is very necessary for we can get a better performance than only one feature even though it is the best one, but too many features are unnecessary for little contribution can be made by these new features. It means feature fusion can help improve the performance of classification but there is an upper bound for adding new features.

4.4. Distribution of Features’ Weights

For every salient object detector, we output the features’ weights after boosting. Simple statistical methods, such as the average and the variance, are used to analyze the importance of each feature. From Table 3, it’s noticed that three color features have much larger weights than four co-occurrence matrix features, and the variances of the distributions of these four co-occurrence matrix features’ weights are very small. That means the color feature is more discriminative than the statistical texture feature when applied to recognize objects on the TV programs. What’s more, the moment feature has a smaller weight than the location feature for the segmentation is very imperfect and we even over-segment the images purposely. Location information is more stable than moments for it only cares about the bounding box of the region, but such information is too little for classifying objects so its weight is still usually below the average level.

5. CONCLUSION

By segmenting and labeling on the key frames of the shots of TRECVID 2005 development video, we define 21 salient objects to detect, which can help achieving a middle-level understanding of the video shots. 11 regional features on color, shape, edge and texture, are extracted. And a boosting method is adopted for feature selection after the SVM training on each feature. Except for a fast convergence speed, our boosting method performs better than two other regular combination methods. And it can be used for selecting the necessary features to reduce the complexity of the prediction.

More features won’t lead to higher performance. So exploring the upper bound theoretically is interesting. Besides,
for a number of salient objects are labeled, the relationship between these objects, such as the ratio of the border length between two regions, to the perimeter of one of these two regions [4], hasn’t been considered which may probably improve the performances of the detectors. And using these detectors to learn the scene-level concepts are also the work we plan to do next step. Furthermore, the auditory salient object is useful especially in videos for the audio supplies some important cues for semantics learning. The development of auditory salient objects detectors is also challenging.

6. References


