TRECVID 2008 High-Level Feature Extraction By MCG-ICT-CAS*

Sheng Tang, Jin-Tao Li, Ming Li, Cheng Xie, Yi-Zhi Liu, Kun Tao, Shao-Xi Xu

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China

{ts, jtli, mli, xiecheng, liuyizhi, ktao, xushaoxi}@ict.ac.cn

ABSTRACT

For TRECVID 2008 concept detection task, we principally focus on: (1) Early fusion of texture, edge and color features TECM, abbreviation of the combined TF*IDF weights based on SIFT features, Edge Histogram, and Color Moments. (2) To improve the training efficiency and explore the knowledge between concepts or hidden sub-domains more easily and efficiently, we propose a novel method based on Latent Dirichlet Allocation (LDA): LDA-based multiple-SVM (LDASVM). We first use LDA to cluster all the keyframes into topics according to the maximum element of the topic-simplex representation vector (TRV) of each keyframe. Then, we train the annotated data in each topic for each concept. During training, unlike multi-bag SVM, we only use positive samples in current topic for the sake of retaining sample's separability, instead of all positive samples among the whole training set, and ignore the topics with too few positive samples. While testing a keyframe for a given concept, we adopt TRV as the weight vector, instead of equal weighting strategy, to combine the SVM outputs of topic-models. (3) Introduction of Pseudo Relevance Feedback (PRF) into our concept detection system for the purpose of making re-trained models more adaptive to the test data: unlike existing PRF techniques in text and video retrieval, we propose a preliminary strategy to explore the visual features of positive training samples to improve the quality of pseudo positive samples. Experimental results demonstrate that our proposed LDASVM approach is both effective and efficient.

Keywords

Concept detection, Latent Dirichlet Allocation, Early fusion, Pseudo Relevance Feedback

1. Introduction

With the rapid growth of multimedia application technology and network technology, processing and distribution of digital videos become much easier and faster. However, in searching through such large-scale video databases, indexation based on low-level features like color and texture, often fails to meet the user's need which is expressed through semantic concepts due to the "semantic gap"[1]. Consequently, how to establish the mapping between the low-level features and high-level semantic descriptions of video content to bridge up the "semantic gap" efficiently, i.e., automatic annotation of video at the semantic level, is currently becoming an important topic in the multimedia research community.

For large-scale video database, training is very time-consuming. Furthermore, it generally consists of many sub-domain data sets with their own characteristics, training a model by mixing data of different sub-domains may lose important information and degrade the performance of the systems [2]. Recently, G. Wang [2] proposed to explore knowledge of sub-domain for concept detection in news videos. However, for other videos such as recent TRECVID videos, since the video database is so large that even sub-domain itself may not be obvious, the knowledge of sub-domain can not be easily explored.

To improve the training efficiency and explore the knowledge between concepts or hidden sub-domains more easily and efficiently, we propose ***LDASVM*** through latent semantic analysis.

The rest of the paper is organized as followed. The overall system, especially our main focuses, is described in Section 2. Then, annotation of training data is introduced in Section 3. Feature extraction, LDASVM and introduction of pseudo relevance feedback in to our concept detection system are described in Section 4 to 6 respectively. We also try another concept detection method called Localization Classifiers in Section 7. Finally, we give our experimental results in Section 8 and draw our conclusion in Section 9.

* This work was supported by National Basic Research Program of China (973 Program, 2007CB311100), National Nature Science Foundation of China (60873165), National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), and the Beijing New Star Project on Science & Technology (2007B071).

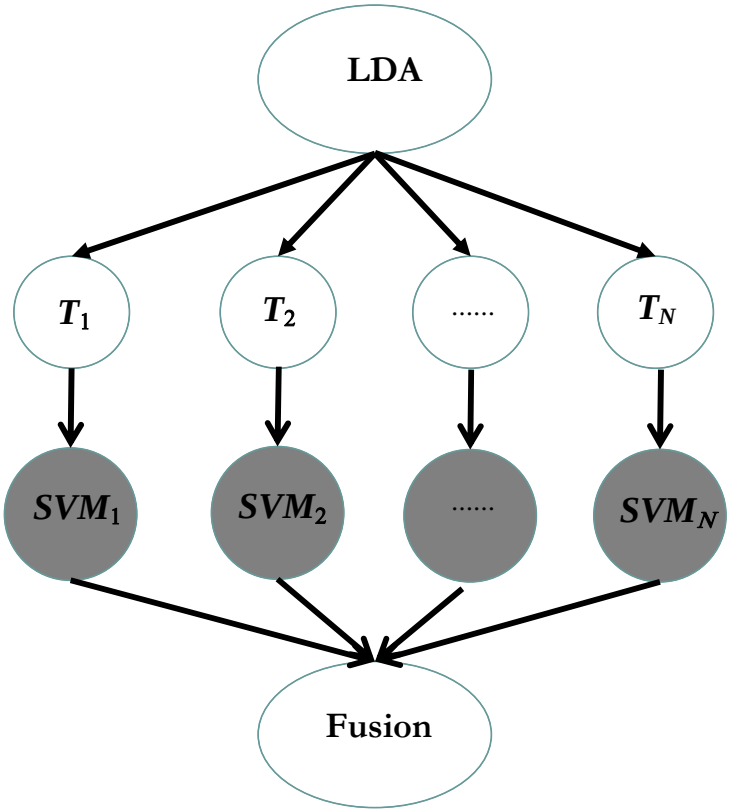

Fig.1 The procedure of our proposed LDA-based multiple-SVM (LDASVM)

2. System overview

For concept detection task, we principally focus on:

(1) ***Early fusion of texture, edge and color features TECM*** (890 dims), abbreviation of the combined TF*IDF weights based on SIFT features (345 dims), Edge Histogram (320 dims), and Color Moments (225 dims).

(2) To improve the training efficiency and explore the knowledge between concepts or hidden sub-domains more easily and efficiently, as shown in Fig.1, we propose a novel semantic concept detection method based on latent semantic analysis: ***LDA-based multiple-SVM (LDASVM)***. We first use LDA to cluster all the keyframes into 20 topics according to the maximum element of the TRV of each keyframe. Then, we train the annotated data in each topic to get SVM model for a given concept. Unlike multi-bag SVM, during training, we only use positive samples in current topic for the sake of retaining sample's separability, instead of all positive samples among the whole training set, and ignore the topics with too few positive samples. While testing a keyframe for a given concept, we adopt TRV as the weight vector, instead of equal weighting strategy, to fuse the SVM outputs of topic-models.

Since the topic size is greatly smaller than the total number of samples and the samples in each topic are of higher separability after latent semantic analysis, the SVM training is very efficient. Moreover, employing all samples in each topic for cross-validation becomes very practicable.

(3) ***Introduction of Pseudo Relevance Feedback (PRF)*** into our concept detection system for the purpose of expanding positive training samples: unlike existing PRF techniques in text and video retrieval, we propose a preliminary strategy to explore the visual features of positive training samples to improve the quality of pseudo positive samples, ***hence making re-trained models more adaptive to the test data***.

(4) We attempt to use a method named Localization Classifiers to improve the classification performance.

(5) Object-based features: we train models with object-based TF*IDF features within labeled rectangles for positive training samples. But the result is not good due to unavailability of such object-based features of test samples.

(6) Annotation of training data: Since our annotation of the 2007 training data was widely used by many participants (65 out of 110 runs used ours) [3], this year we kept on providing full annotation of the 2008 training data.

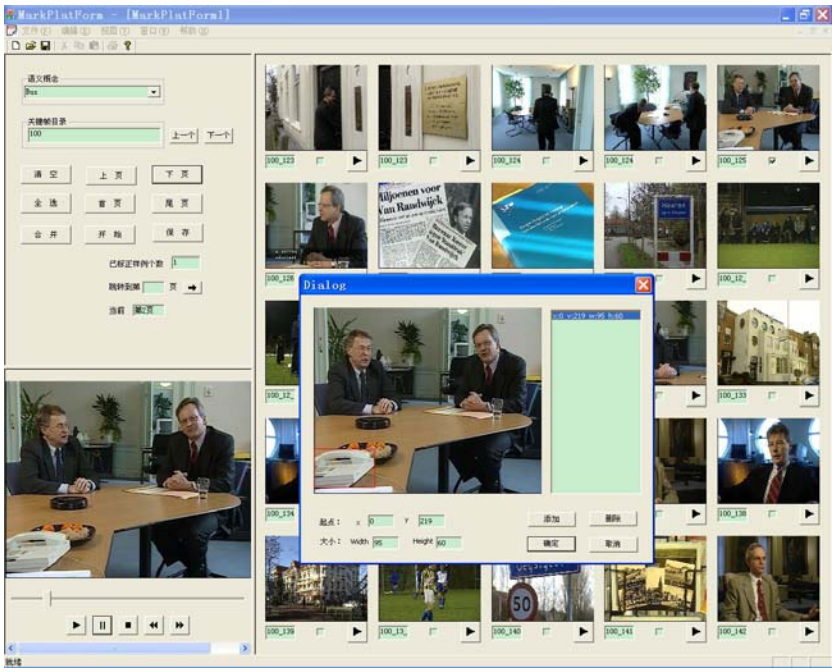

Fig.2 The interface for our annotation of training data

3. Annotation of training data

In TRECVID 2007, we provided full annotation of the 2007 training data for the multimedia research community. We organized 16 persons in our group to annotate 36 concepts of the 21532 TRECVID 2007 development keyframes extracted by George Quenot et al [4], and each concept was annotated by only one person.

In TRECVID 2008, we organized 15 persons to annotate 20 concepts of the 39674 TRECVID 2008 development keyframes (including 21532 TRECVID 2007 development keyframes and 18142 TRECVID 2007 test middle I-frames extracted by us). Each concept was annotated by one person, and checked by another person. Therefore, we arranged 10 persons to annotate 20 concepts, and another 5 persons to check all the annotation.

In order to encourage researchers to propose methods extracting features based on object rather than the whole frame, we divided the 20 concepts into two groups:

(1) Object-related concepts: we located and recorded the rectangle range of the local object such as the telephone shown in the Fig.2. This group includes 14 concepts:

002 Bridge, 003 Emergency_Vehicle, 004 Dog, 006 Airplane_flying, 007 Two people, 008 Bus, 009 Driver, 012 Telephone, 014 Demonstration_Or_Protest, 015 Hand, 016 Mountain, 018 Boat_Ship, 019 Flower, 020 Singing.

(2) Scene-related concepts: we regarded the whole frame as an object concerned with the concepts. This group includes 6 concepts:

001 Classroom, 005 Kitchen, 010 Cityscape, 011 Harbor, 013 Street, 017 Nighttime.

We attempted to provide more information about object, so we selected more concepts to the object-related concepts, such as 014 Demonstration_Or_Protest, 020 Singing for locating the relative persons.

It took about 40-45 hours to annotate each concept of the first group (including drawing the boxes), and about 15-20 hours for the second group.

4. Feature Extraction

We extract six basic visual features [5] for each key frame of the video shots. The basic visual features are:

(1) Color Histogram (CH): 166 dims;

(2) Color Correlogram (CC): 166 dims;

(3) Color Moments (CM): 225 dims;

(4) Co-occurrence Texture (CT): 96 dims;

(5) Wavelet Texture Grid (WTG): 108 dims;

(6) Edge Histogram (EH): 320 dims;

We build a visual vocabulary of SIFT points from keyframes. We choose approximately 3,000 keyframes from TRECVID 2008 development set, containing only

positive samples over all the 20 concepts. However, these keyframes have already included all kinds of visual information of negative samples because of the intra class diversity in the development set. With K-mean clustering, 981,231 SIFT points (for keyframes with object-related concepts, we only extract SIFT points within the range of labeled rectangles) are quantized into 345 clusters, and each cluster represents a visual keyword. As depicted in [6], we compute:

(7) TF*IDF weights based on SIFT features:345 dims.

To study the early fusion technique, we combine texture, edge and color features:

(8) TECM (890 dims): (7)+(6)+(3).

5. LDASVM

To improve the training efficiency of concept models and to exploit the relationship between concepts, we propose a novel method: ***LDASVM*** which includes LDA clustering, SVM Training, and TRV-weight-based fusion.

5.1 LDA clustering

After quantization of the TF*IDF weights, we use Latent Dirichlet Allocation to cluster all the keyframes into 20 topics according to the maximum element of the topic-simplex representation vector (***TRV***) of each keyframe.

5.2 SVM Training

We train SVM models for all the 20 topics and 20 concepts. Unlike multi-bag SVM, during training, we only use positive samples in current topic for the sake of retaining sample's separability, instead of all positive samples among the whole training set.

If the clustered training topic is too small or unbalanced, we may not have sufficient positive training data. Therefore, we ignore the topics with too few positive samples. In this year's task, for all the 20 concepts, we get 344 models after removing 56 topics with no more than 1 positive sample.

5.3 TRV-weight-based fusion strategy

While testing a keyframe for a given concept, we adopt TRV as the weight vector, instead of equal weighting strategy as usually adopted by multi-bag SVM method, to fuse the SVM outputs of topic-models.

In our experiments, we standardize TECM feature matrix by removing the mean of each column and dividing each column by its standard deviation when training SVM Models, and use the mean and standard deviation to standardize TECM feature of test key frame.

6. Pseudo Relevance Feedback

With past years of research, Pseudo Relevance Feedback (PRF) has shown its great potential in information retrieval. Its basic idea is to extract query expansion examples from the top-ranked retrieval results to formulate a new query for a second round retrieval, and its effectiveness strongly relies on the quality of selected expansion examples.

However, it has seldom been introduced to the high level feature detection task. Perhaps it may be a distinctive advantage that there are much more positive training examples available in concept detection than in information retrieval. For the purpose of expanding positive training samples and making re-trained models more adaptive to the test data: unlike existing PRF techniques in text and video retrieval, we propose two preliminary strategies to explore the visual features of positive training samples to improve the quality of pseudo positive samples. One is similarity-based and the other is detector-based.

As for similarity-based method, we select pseudo positive samples by calculating the feature similarities between top-retrieval examples with positive training samples after every retrieval process. While for detector-based method, we select pseudo positive samples through the overall evaluation of positions among the ranked lists from several detectors.

7. Localization Classifiers

An object or scene corresponding to a semantic concept shows variant appearance in visual or other aspects, and an action or behavior is even more difficult to be described. So the samples are often scattered irregularly in a feature space, but no general simply model can be used to describe such distribution. Here we attempted to use a method named Localization Classifiers to improve the classification performance. We segment feature space into many sub-regions, investigate the sample distribution in sub-regions, and train the sub-classifiers for sub-regions if necessary. If the sub-regions are segmented in suitable scale, they can preserve enough information of total sample distribution. At the same time, as we only need to investigate a little part of the samples for each time, we can see that its distribution is simpler than global distribution, sometimes approximate to linearly separable, which is much more favorable for classification.

To accomplish the segmentation of sub-regions, we first investigate the sample density and training sample number in order to decide the proper number of sub-regions. Then we used K-means clustering on training data, and each cluster is regarded as a sub-region. By counting the positive and negative sample numbers, the prior probabilities of positive samples are calculated for all sub-regions. A

sub-regions with enough number of samples and enough large prior probability is named as significant sub-region, others are insignificant. Then we only train SVM sub-classifier for significant sub-regions respectively.

A testing sample first should be assigned to a nearest sub-region. If the sub-region is insignificant, its prior probability will be used as an approximation of the probability that the sample is positive. If the sub-region is significant, the final output probability is decided by both the prior probability and the output of SVM sub-classifier, according to the following equation:

$$P_{output} = P_{prior}^{(\log_2 \frac{1}{P_{SVM}})}$$

As the complex global distribution model resolving is decomposed into resolving some simple distribution models, the performance of Localization Classifiers is obviously superior to traditional classifiers. On the other hand, the training time is also saved as a big train sample set is divided into some small sets.

8. Experiment

We submitted a total of 6 runs. The description and MAP of each run are shown in the following Table 1.

Table 1 Description and MAP of our HLF runs

HLF run	MAP	Description
A_ICT_1	0.048	Visual Baseline
A_ICT_2	0.038	LocalizationClassifier
A_ICT_3	0.065	TECM_LDA_SVM
A_ICT_4	0.037	TECM_LDA_SVM_PRF
A_ICT_5	0.076	TECM_LDA_SVM+Baseline
A_ICT_6	0.078	Fusion All

Our run A_ICT_3 (mean infAP 0.065) shows that TECM-feature-based LDASVM method is very effective compared with our baseline A_ICT_1 (0.048) with 35.4% improvement. Since the topic size is greatly smaller than the total number of samples and the samples in each topic are of higher separability after latent semantic analysis, the SVM training is very efficient, only about twenty minutes for all the 344 models on our server. Moreover, employing all samples in each topic for cross-validation becomes very practicable (generally less than 2 hours for each model in our experiments).

Although we achieve good results especially for concepts with very few positive training samples in our experiments, our submitted run A_ICT_4 only got mean infAP of 0.037, much lower than A_ICT_3. It shows that our PRF method is not stable since the introduction of pseudo positive samples may ruin the separability of topic samples.

The infAP of A_ICT_2 is not as high as we expected, it might be caused by two reasons: first, if a testing sample is far from any original sub-region, or falls into a sub-region with little training samples, the predict result will become unreliable, that is, the generalization performance needs to be improved; second, we didn't train the C and gamma parameters for the sub-classifiers, but the results show that this step can not be neglected. However, the method still show some individual highlights in several concepts.

To investigate the effectiveness of object-based features, we also train models with TF*IDF features within labeled object rectangles for positive training samples. But we can not extract such object-based features for the test data since we know nothing about whether an object is in a keyframe. Hence the result is not as good as that of training models with TF*IDF features of the whole keyframe.

We only use one keyframe per shot for the test data, and our infAPs should be further improved if we use more keyframes per shot.

9. Conclusion

In summary, the early fusion TECM feature, clustering via LDA, sample's separability-keeping strategy during training and TRV-weight-based fusion strategy during testing together contribute to the high efficiency and effectiveness of our proposed method. On the other hand, the determination method of hidden topic number should be carefully studied for further improvement.

Additionally, how to utilize positive training samples to select more pseudo positive samples as much as correctly and filter the pseudo positive samples which may destroy the topic's separability, should be further studied.

REFERENCES

[1] Hauptmann, A., Yan, etal; "Filling the Semantic Gap in Video Retrieval: An Exploration"; Semantic Multimedia and Ontologies, page(s):253-278, 2008.

[2] G. Wang, T.-S. Chua, M. Zhao, "Exploring Knowledge of Sub-domain in a Multi-resolution Bootstrapping Framework for Concept Detection in News Video", to appear in ACM Multimedia 2008

[3] W. Kraaij, P. Over, G. Awad, "TRECVID-2007 High-Level Feature task: Overview", TRECVID 2007 Workshop, Gaithersburg, MD, USA, November 5-6, 2007.

[4] Stéphane Ayache and Georges Quénot, "TRECVID 2007 Collaborative Annotation using Active Learning", TRECVID 2007 Workshop, Gaithersburg, MD, USA, Nov., 2007.

[5] Arnon Amir, Janne Argillander, Murray Campbell, etal, "IBM Research TRECVID-2005 Video Retrieval System", TRECVID 2005 Workshop, Gaithersburg, MD, Nov. 2005.

[6] Josef Sivic and Andrew Zisserman, "Video Google: a text retrieval approach to object matching in videos"; In Proc. ICCV, Oct 2003.

TRECVID 2008 Search Task by MCG-ICT-CAS*

Juan Cao, Yong-Dong Zhang, Bai-Lan Feng, Xiu-Feng Hua, Lei Bao, Xu Zhang and Jin-Tao Li

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China

{caojuan, zhyd, fengbailan, huaxiufeng ,baolei, zhangxu, jtli }@ict.ac.cn

ABSTRACT

This paper describes the MCG-ICT-CAS automatic search system for TRECVID2008. In the concept-based module, we proposed a novel distribution based concept selection (DBCS) approach, which achieved a stable good performance for all the topics (0.053). In the visual-based module, we focused on the low dimensional semantic features by Latent Dirichlet Allocation model and get an infAP of 0.033. Finally, a re-ranking technology based on the motion and face and a multi-runs and multi-examples fusion approach (SSC) were applied to aggregate the basic search results, which produced a significant improvement.

Keywords

Distribution based concept selection (DBCS), Latent Dirichlet Allocation (LDA), SSC fusion re-ranking

1. Introduction

In automatic search, we focused on concept-based retrieval by DBCS and the visual retrieval based on the low dimensional feature in LDA semantic space. Moreover, we extend our emphases to the dynamic fusion and re-ranking methods. The framework of our automatic search system is shown in Figure 1, and the runs we submitted are:

F_A_1_MCG-ICT-CAS_1: re-ranking results of SSC dynamic fusion of Run_2 and Run_4

F_A_2_MCG-ICT-CAS_2: SSC dynamic fusion of HLF results by the Distribution Based Concept Selection(DBCS) method and HLF baseline(Run_5)

F_A_2_MCG-ICT-CAS_3: re-ranking results based on Run_5 by face and motion information

F_A_2_MCG-ICT-CAS_4: visual baseline by Latent Dirichlet Allocation (LDA) and multi-bag SVM

F_A_2_MCG-ICT-CAS_5: high-level-feature(HLF) baseline by multi-bag SVM retrieval

F_A_2_MCG-ICT-CAS_6: text baseline by lucene retrieval.

Table.1 the performance of six runs for automatic search

Run_ID	Mean InfAP
F_A_1_MCG-ICT-CAS_1	0.067
F_A_2_MCG-ICT-CAS_2	0.053
F_A_2_MCG-ICT-CAS_3	0.036
F_A_2_MCG-ICT-CAS_4	0.033
F_A_2_MCG-ICT-CAS_5	0.029
F_A_2_MCG-ICT-CAS_6	0.009

The corresponding performances of these six runs are listed in Table.1. Overall, our contributions are following:

- Distribution based concept selection(DBCS) method. It aims to select the concepts with the most statistical discriminability for the topic by analyzing the difference of the feature distribution between the examples and whole collection. Its stable good performance in all the topics makes the greatest contribution to the highest mean infAP(0.067).
- The sift-visual-keywords feature in the low dimensional LDA semantic space. We adaptively find the top-k(950) important sift key-points clusters in all 48 query visual examples by a density-based K-means method, and further reduce it into the 80 dimensional semantic space by LDA. This feature obviously outperforms the other features, and achieves a mean infAP of 0.028.
- Re-ranking based on the motion and face. We extract the shot-level semantic motion and face features, and accomplish our experiments of re-ranking algorithm based on Run_5. 27 topics among 48 are automatically judged as motion-related or face-related by the algorithm, and the performances of 23 among them are improved. Run_3 had the mean infAP of 0.036 with 24% improvement than that of the Run_5.

* This work was supported by National Basic Research Program of China (973 Program, 2007CB311100)，National Nature Science Foundation of China (60873165), National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), and the Beijing New Star Project on Science & Technology (2007B071).

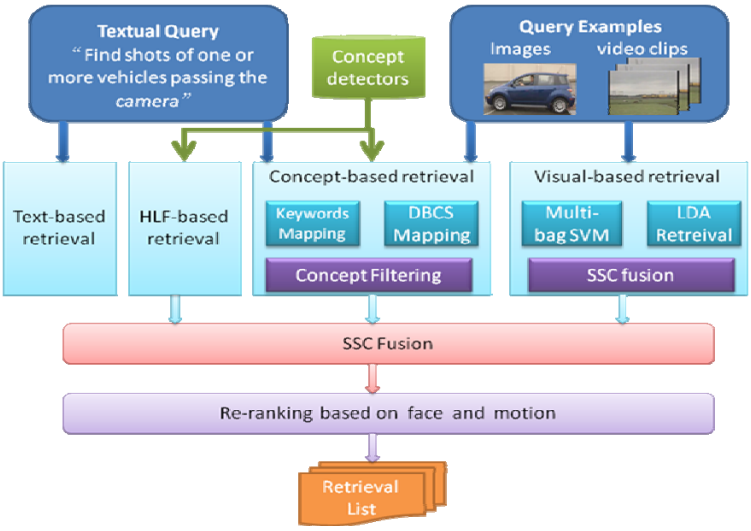

Fig. 1 MCG-ICT-CAS automatic search framework

- Dynamic fusion method based on the Smoothed Similarity Cluster(SSC) method. We implement all the fusions in our system by SSC, and more than 80% of the fusions improved the original results. The improvement is bigger when the feature difference between the original results is greater.

2. Text-based Retrieval

The text-based run consists of pre-processing and retrieval parts. Firstly we apply pre-processing for textual queries and ASR texts of every shot in the test data set, including stemming and stop words removal. Secondly, we use the lucence[1] to build index and implement retrieval.

3. HLF-based Retrieval

In the high-level-feature based retrieval, we regard the concept detector scores as the high level feature, and learn the SVMs for every topic on it. This year we use the concept detectors results of CU-VIREO374 [2], and employ a multiple modeling Support Vector Machine method, named Multi-bag SVM[3] as the basic classifier. For every topic, 10 SVMs are trained, where the positive examples are the topic examples[1] and the negative ones are randomly sampled from test data set without repetition. We used RBF kernels in SVM, which have two primary parameters: C and γ. To solve the efficiency problem of the traditional cross validation(CV) in the parameters selection, we elide the process of 'cross' and randomly divide the bag into train set (70%) and validation set (30%). All the possible parameters are used in training and validating, and the pair which maximizes the precision is chosen as the parameters of the bag. In our experiments of TRECVID07 dataset, the parameters selected by this method outperform the traditional CV. Furthermore, this method can reduce the time consumption of modeling and classification to about one minute per topic.

[1] All the "topic examples" in this paper include the given image examples and the keyframes extracted by our clustering-based method from the query video clips

4. Visual-based Retrieval

Last year, we applied Latent Dirichlet allocation (LDA) [4] to the search task. This year we try to combine the LDA and Multi-bag SVM to improve the visual-based retrieval performance. LDA is a statistical topic model which can find out the low dimensional latent semantic space in the test corpus. The features representing in this space are more semantic and more robust than all the original low-level visual features space in our experiments conducting on TRECVID06 and 07 datasets. Furthermore, since the learning of LDA model is based on the test corpus but not limited to the query visual examples, its performance is more stable to the number of visual examples than the multi-bag SVM. The details are following:

Step1. Visual Feature Extraction

The visual feature we used this year includes the block edge histograms (EH), block color moments (CM), Sift-visual-keywords, their low dimensional semantic features reduced by LDA, and their early fusion features. The sift-visual-keywords is the cluster centers of all the sift key-points in all 48 query visual examples by a density-based adaptive K-means method [5].

Step2. The Multi-bag SVM and LDA Retrieval

We train the Multi-bag SVMs for every topic based on the features in step1, and obtain a group of 48 relevant lists for each feature.

For each feature, we build a LDA model in test corpus and obtain the inference of the data in test corpus and the query visual examples. Then, we compute the similarity

between the test data and queries in the latent semantic space. As there are several examples for each query, we fuse the relevant lists of multi-example by SSC fusion method.

Step3. Fusion

For each feature, we dynamically fuse the relevant lists of the multi-bag SVM retrieval and LDA-based retrieval results by SSC.

5. Concept-based Retrieval

Concept-based retrieval method has played a key factor to improve the performance in many automatic video retrieval system[8] [11]. It applies the results from off-line concept detection and the on-line query-to-concept map. Natsev et al. comprehensively summarized the mapping technologies in concept-based retrieval so far into semantic similarity principle and the statistical relevant methodology[11]. The nature target of these methods is finding the most similar concepts for the query. Differently, we select the most discriminative concepts to distinguish query examples and test collection, and provide an effective concept selection method based on the concept distribution(***DBCS***).

In addition, we also select the concepts by keywords as the supplement for DBCS. We firstly extract keywords from the textual description of the queries, including the nouns and verbs. Then we map the keywords to concept name directly. After combining the two concept mapping results, we averagely fuse these detector scores as the final concept-based retrieval results.

Distribution Based Concept Selection (DBCS)

Symbols:

t: a concept.
V: the collection of the concept.
s: a shot.
$\{c_1,c_2,\ldots,c_m\}$: the category set.
n_i: the number of shots belonging to c_i.
F(t, s): the distribution function of t in s.
$F(t, c_i)$: the distribution function of t in c_i.

Definition 1 $VAC(t,c_i)$ measures the difference between distribution of t in c_i and distributions of t in other categories.

Definition 2 $VIC(t,c_i)$ is the distribution difference of t in all shots belonging to c_i, measured by the variance of F(t,s) where $s \in c_i$.

For a given category, the distributions of its features should fluctuate widely between this category and other categories, but remain stable within this category. So features representing category c_i should have large $VAC(t,c_i)$ but small $VIC(t,c_i)$. DBFS method aims at finding out this kind of feature. In our case, the question is degenerated into a two categories issue. One is the query category, the other is the test collection.

The algorithm of selecting features for a topic based on DBFS can be described as follows:

Algorithm: the description distribution based feature selection(DBFS) algorithm

1: For each t in V, For each s in C:
$F(t,s) \leftarrow P(t \mid s)$

2: For each c_i in $\{c_1,c_2,\ldots,c_m\}$:

$$F(t,C_i) \neg \frac{1}{n_i}\sum_{s\in c_i} F(t,s)$$

$$VAC(t,c_i) \leftarrow \sum_{j\neq i} sign(F(t,c_i)-F(t,c_j))(F(t,c_i)-F(t,c_j))^2$$

$$VIC(t,c_i) \leftarrow \frac{1}{n_i}\sum_{s\in c_i}(F(t,s)-F(t,c_i))^2$$

3: for each t:
$Score(t) = VAC(t,c_i)/VIC(t,c_i)$

4: sort features by score in descending order and select first k features

6. Re-ranking based on the Motion and Face Information

6.1 Motion

This year we extract the motion features from two aspects: one is the global camera motion patterns; the other is the foreground motion information.

6.1.1 Camera motion

We extract a 8 dimensional camera motion feature from motion vectors of p-frames in compressed domain, including tilt up, tilt down, pan left, pan right, zoom in, zoom out, still and unknown[6].

6.1.2 Foreground motion

We extracted a semantic-level motion feature based on two low-level motion features, which can covers spatial and temporal characteristics simultaneously. The re-ranking algorithm consists of three main stages:

Frame-level motion feature extraction: We extract a 18 dimensional frame-level motion feature from the motion vectors[7], which includes foreground moving region' area, the coordinates of centroid and 14 invariant moments information. This feature clearly reveals the size, location, motion direction, and motion intensity of moving region obviously.

Shot-level motion feature extraction: We extract a 38 dimensional shot-level motion feature by computing the variances and means of the frame-level motion features of the consecutive frame sequence in one shot.

Semantic-level motion feature representation: We filter the potential noisy frames based on the above two

levels motion features, such as the marginal frames and the ones including disordered small moving regions. Then we link the centroid of the distinct moving region's across frames to form the foreground moving region trajectory, and further quantize the trajectory to 4 direction values. This value unified the information of the semantic content of the event and the general shooting methods for this kind of event. It reflects the statistic semantic characteristic of the moving events, and is more accurate and robust to describe the event.

Then we extract the motion intensity from the shot-level motion feature as the **Motioncoefficient**, and compute the ***MotionScore*** between queries and shots in database based on the Semantic-level motion feature.

6.2 Face

Face is one of the most frequent elements, which appears in 75.72% of the test shots. We extract 2-dimension shot-level face feature for each shot of the query videos and test ones, representing the average face number and the average face size. With this feature, a ***faceScore(FS)*** and ***faceCoefficient(FC)*** are generated for each shot, which are used to re-rank the initial result list. We give the steps as follows.

Step 1: Face feature extraction. Face detection is applied to video shots every 5 frames to get the face size and position of each frame. For the detected frames, the average face number and the average face size for each frame are computed as the face feature of the query and the test shot.

Step 2: FS and FC calculation. The ***faceScore*** measures the similarity of face information between the query and shot, which is calculated as the distance of the scaled face features. Besides, the face information conducts different extent of impacts toward different queries. It is important for topics as "Find shots of a person talking behind a microphone" while is useless for topics as "Find shots of one or more vehicles passing the camera ". Since that, we regarded the average face number of the query examples as the ***faceCoefficient***, which indicates the importance of the face information. As the feature data we extracted, the average face number of Topic0254 is 0.984375 while Topic0230 is 0.

6.3 Re-ranking

We select the motion-related or face-related queries by analyzing the feature distribution in the query examples collection and test corpus, and update the score of the initial rank list separately based on the motion and face information as following re-ranking form:

$$Score' = Score + FactorScore \times FactorCoefficient$$

7. Dynamic Fusion

We employed the score distribution based automatic coefficient generation approach in the fusion of multi-run and multi-example. The main idea of our approach derived from the observation that "if a feature undergoes a rapid change in its normalized scores, then that feature is likely to perform better than a feature which undergoes a more gradual transition in normalized scores" [11]. P. Wilkins use the SC (Similarity Cluster)[11] to measure the performance of the result for a query. Theoretically the big SC values are important for a well performed run, but the value is unstable in the real data. When we compute the mean of the 48 SC values, some outer values can greatly impact the result. So on the base of SC values of 48 queries, we calculate a ***Smoothed Similarity Cluster (SSC)*** value for each run, which determines the coefficient of the run. We take the median of the 48 SC values to measure the relative performance of a run and smoothed with the standard deviation.

The SSC value is computed as follows:

$$SC = \frac{\frac{1}{1000}\sum_{n=1}^{1000}(score(n) - score(n+1))}{\frac{1}{N}\sum_{n=1}^{N}(score(n) - score(n+1))}$$

$$SSC = \frac{median(SC)}{standard\ \ deviation(SC)}$$

And the coefficient of the run is calculated as follows.

$$Run\ Weight = \frac{Run\ SSC\ Score}{\sum All\ SSC\ Scores}$$

In our system, all fusion processes are realized by SSC method.

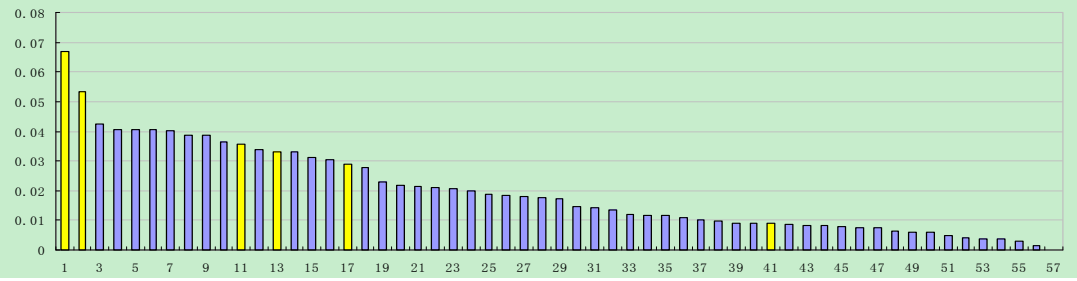

Fig. 2 The performances of six submitted runs for automatic search. The yellow bars are from MCG-ICT-CAS.

Table 2 The comparative analysis of DBCS method and keywords mapping method in concept-based retrieval.

Topic	DBCS method	Keywords method
Topic0261: Find shots of one or more people at a table or desk , with a computer visible	**infAP: 0.116**	**infAP: 0.012**
	1:Attached_Body_Parts 2:Classroom	1:Computers
Topic0231: Find shots of a map	**infAP: 0.008**	**infAP:0.137**
	1:Text_On_Artificial_Background 2:Maps	1:Maps

8. Experiments and Analysis

We submitted 6 automatic type A runs for search task. Figure 2 shows their performances in all the runs.

Our text baseline run(Run_6) had the lowest infAP of 0.009.

Our high-level feature based run(Run_5) used the scores of the concepts detectors as the high-level features, and retrieved based on the multi-bag SVM method with fast parameter selecting. This run produced an infAP of 0.029.

Our visual baseline run(Run_4) fused the LDA and multi-bag SVM results, and had an infAP of 0.033.

Based on the Run_5, we accomplished two experiments. One is the re-ranking algorithm with motion and face features. 27 topics among 48 are judged as motion-related or face-related by the algorithm, and the performances of 23 among them are improved. Run_3 had the infAP of 0.036 with 24% improvement on the HLF-baseline.

The other experiment on Run_5 is the concept-based retrieval with DBCS method. The fusion result produced an infAP of 0.053 with an improvement of 83%. In order to have a deep understanding about DBCS, we evaluated three unsubmitted runs by the ground truth of NIST. The DBCS method achieved a stable good performance in all the topics(mean infAP=0.039). It aims to select the concepts with the most statistical discriminability for the topic. For example, in the topic0261 of Table.2 we can find that, the pictures with the computer visible is always with hands visible, so the DBCS selected the “Attached_Body_Parts” and “classroom”, which are less semantic relevant but most statistical relevant to the query. On the other hand, the keywords method is try to find the semantic relevant concepts. For the limit of the lexicon, it has 7 zero-mapping topics in 48, and has a mean infAP of 0.026. It is not as stable as DBCS, but it can catch the exact semantic of the simple queries and achieve the best performance.

After fusing the visual and concept-based results by SSC method, we get the highest infAP of 0.067. The SSC dynamic fusion method can make improvement in more than 80% cases.

References

[1] J. Lucence, "Jakarta Lucene Text search engine in Java",http://jakarta.apache.org/lucene/docs/index.html

[2] Y.-G. Jiang, A. Yanagawa, S.-F. Chang, and C.-W. Ngo, "CU-VIREO374:Fusing Columbia374 and VIREO374 for Large Scale Semantic Concept Detection", Columbia University ADVENT Technical Report #223-2008-1,Aug. 2008.

[3] J. Tesic, A. Natssev, and J.R. Smith. Cluster-based data modeling for semantic video search. In ACM International Conference on Image and Video Retrieval (ACM CIVR), 2007.

[4] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993–1022, January 2003

[5] XIE Yu-xiang, WU Ling-da, LAO Song-yang, WANG Chen, HU Ya-nan, An efficent indexing algorithm of clustering supporting QBE image retrieval for large image databases Mini-micro system, 2001

[6] X. Zhu, A.K. Elmagarmid, X. Xue, L. Wu, and A.C. Catlin, “InsightVideo: Toward Hierarchical Video Content Organization for Efficient Browing, Simmarization and Retrieval”, *IEEE Transactions on Multimedia*, Aug.2005.

[7] Tao Kun, Wu Si, Lin Shouxun, Zhang Yongdong, “Research on Panorama Composition Technique of Sports Video”, Journal of computer-aided design and computer graphics, Nov. 2005.

[8] M. Campbell, A. Haubold, M. Liu, A. P. Natsev, J. R. Smith, J. Tešić, L. Xie, R. Yan and J. Yang，IBM Research TRECVID-2007 Video Retrieval System， In NIST TRECVID Video Retrieval Workshop. 2007.

[9] T. Mei , X. Hua , W. Lai , L. Yang Z. Zha, Y. Liu, Z. Gu, G. Qi, M. Wang, J. Tang, X. Yuan ,Z. Lu and J. Liu, MSRA-USTC-SJTU AT TRECVID 2007: HIGH-LEVEL FEATURE EXTRACTION AND SEARCH, In NIST TRECVID Video Retrieval Workshop. 2007.

[10] Tat-Seng Chua, Shi-Yong Neo, etal. TRECVID 2007 Search Tasks by NUS-ICT.

[11] A. Natsev, A. Haubold, J. Te˘si´c, L. Xie, and R. Yan. Semantic concept-based query expansion and re-ranking for multimedia retrieval: A comparative review and new approaches. In ACM Multimedia (ACM MM), Sep. 2007.

[12] Peter Wilkins, Tomasz Adamek, Gareth J.F. Jones, Noel E. O’Connor and Alan F. Smeaton TRECVid 2007 Experiments at Dublin City University

TRECVID 2008 Content-Based Copy Detection By MCG-ICT-CAS*

Yongdong Zhang, Ke Gao, Sheng Tang, Xiao Wu, Xiaoyuan Cao, Huamin Ren,Yufen Wu, Jian Yang

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China

{zhyd, kegao,ts,wuxiao,caoxiaoyuan,renhuamin,wuyufeng,yangjian}@ict.ac.cn

ABSTRACT

We participated in the content-based copy detection task in TRECVID 2008. This paper describes the details of our system for the task. We propose a fusion method which integrates four modules corresponds to different kinds of transformations separately. For global transformation, this paper presents a feature descriptor called Block Gradient Histogram. Harris-based local descriptor with spatial neighborhood information is proposed for partial occlusion and shift. We also explore a method for Picture in Picture (PIP) transformation particularly, which composed of edge detection and frame localization of PIP. To deal with flip (vertical mirroring), a rotate-invariant descriptor is proposed. With the matching result of keyframes extracted based on video shot boundary detection, the Time Sequence Consistency method is used to improve the precision of video copy detection and time orientation. Experiments show that our methods are efficient and effective.

Keywords

Content-based copy detection, Hierarchical fusion method, Block Gradient Histogram, spatial neighborhood feature

1. System overview

Considering the transformations used for generating the video queries for the copy detection pilot task, we divide all the transformations into 4 types based on their primary transformations, and design a module separately for each of them:

(1) Global quality decrease such as blur, adding noise, change of gamma、resolution and contrast, etc.

(2) Partial content alteration such as occlusion, shift, crop, and insertions of pattern (including the Picture in Picture type 2, the original video is in the background).

(3) Picture in picture type 1(The original video is inserted in front of a background video).

(4) Flip (vertical mirroring).

The quality of copy detection not only depends on the type of transformations, but also on the property of query segments. Because most of the query segments are very short, and do not have much motion and activity, the motion information are not used in our system. A candidate video sequence is defined as a set of successive keyframes described by features (global features or local features). We use these features to progress an approximate search [2] in the database and get the similarity of keyframe-pairs. Afterwards, the Time Sequence Consistency method is used to locate the copies' boundaries and find their temporal position. At last, the detection result of each module is fused to make the final decision. The flowchart of our system is shown as figure 1.

2. Block Gradient Histogram

Frame Features Extraction

Our system extracts Block Gradient Histogram features for each key frame of the video shots. Motivation of this is to rapidly filter out candidate copies which are similar to query video in global frame appearance. Here we extract DC coefficients of intra frames from MPG1 video, and divide each keyframe into 9 blocks. Then global gradient histogram of frames (8*9=72 dimensions) is extracted as visual feature (ref. to Figure 2). The procedure need not decode the video and greatly accelerates the processing. We also eliminate letter box in frames to improve the feature discrimination power. The block gradient histogram based feature is invariant to certain transformations, e.g. lighting, color and global changes. Using all the frames and conducting frame-to-frame matching could guarantee the accuracy of copy location compared to keyframe-based approaches.

*This work was supported by National Basic Research Program of China (973 Program, 2007CB311100), and National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), and National Nature Science Foundation of China (6073056,60873165).

Feature Indexing Structure

The ANN indexing structure is adopted in our framework for feature storage and search. First ANN algorithm builds database to store the large scale frame corpus. Then we use rapid Q-Range Nearest Neighbor search (NRNN) to query the database, which returns all reference frame features in feature space whose distance to the query is smaller than Q. The expected time for a search is logarithmic in the number of elements store in the database.

Time Sequence Consistency method for Optimal Copy Location

By searching database using n query frame feature, n result lists are obtained and each list contains L_i (i=0, …, α) reference frames. α is fixed to be 500 to limit computation cost. Using these results as nodes and the feature similarity as node weight, we construct a graph for copy location. Edges are added if temporal constraints are satisfied: 1) frames f_a and f_b belong to the same reference video; 2) time stamp difference $0 <= T_b - T_a <= \mu$. μ is empirically set as 10 (sec). A copy will be located if total weight of the maximum flow in graph is larger than β (fixed to 20 empirically) as in formula (1) (2).

$$M = \max_{i,j,X} \sum_{l=i}^{j} weight(node_l) \tag{1}$$

$$location = \begin{cases} frame_i \ \ldots \ frame_j \ \ of \ \ videoX, \ \ if \ \ M > \beta \\ none \ \ \ \ else \end{cases} \tag{2}$$

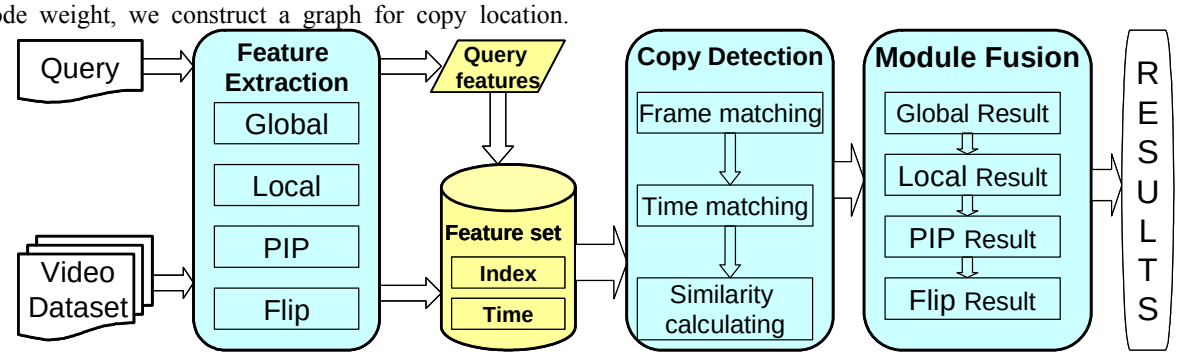

Figure 1: Flowchart of our CBCD system

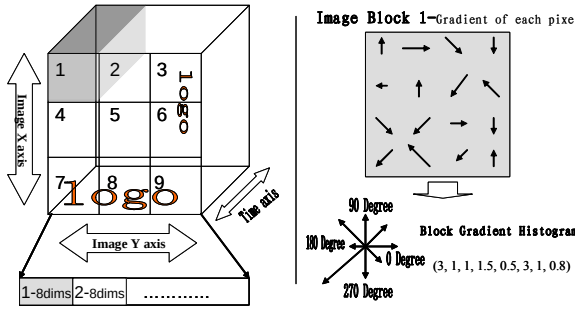

Figure 2: Illustration of Block Gradient Histogram

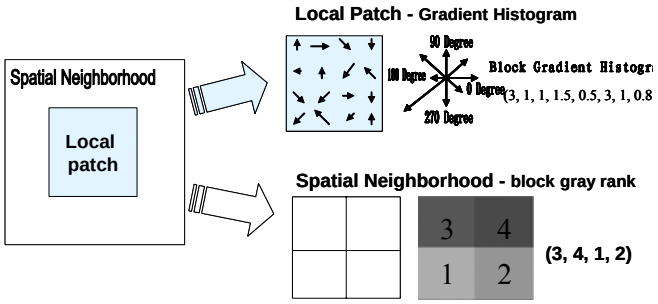

Figure 3: Illustration of local feature with spatial information

3. Local Features with Spatial Information

To deal with partial content alteration just as occlusion and crop, we use harris algorithm [3] to detect salient local point and extract 8-dims gradient histogram from each local patch (as shown in part2). In order to reduce the illegibility of local feature, we present a method to describe its spatial information, as Fig 3. The neighborhood of local point is divided into 4 blocks (scale is 2 times of the local patch), and these blocks are sorted using their average gray level. The rank of each block is used as the spatial signature of each local point. With this information, we can effectively reduce most false matching of local feature.

4. PIP Copy Detection

Picture in picture type1 (PIP) is the original video been inserted in a front of a background. As the smaller video only takes little part of the whole image, i.e. only 20 percent of the original image, common methods can't perform well under this situation. Therefore, we design a module for this kind of transformation separately: adaptively locate the PIP boundary based on edge detection and fusion method, as shown in Fig 4. And finally, Block Gradient Histogram method for Global transformation is used for PIP copy location.

Firstly, we get the key-frames based on shot boundary detection. Secondly, multi-scale and adaptive canny algorithm [4, 5] is used to detect edges for each key-frame. Thirdly, we fuse these results to get a fusion boundary of PIP. Fourthly, with Probabilistic Hough Line Detection, we get all the potential line segments for the PIP region. Finally, we use rule-based method to determine the PIP borders. After this, the detection result is returned.

5. Flip Features for Vertical Mirroring

As to the vertical mirroring, we use a very simple method to deal with it. Because there is only difference in left and right position, so we only change some dimensions of the local and global feature to simulate flipping, and the other work is almost the same.

6. Result Fusion

We tried number of fusion method including non-hierarchical method and hierarchical method. Non-hierarchical method means we use 4 modules to calculate each query separately at the same time, and only the result with the maximal score will be submitted. It is very simple but not efficient enough, because the global feature based method is very fast and effective, so we can use it to filter some result. Accordingly, the hierarchical method submits the result of each module in turn. For each query video, if any previous module has found its corresponding video clip, then we will submit the result, and go on to calculate the next query. Tested in the develop set, hierarchical method performs the best for most test queries, and the processing time is reduced greatly.

7. Experiments and Result Analysis

We submitted a total of 3 runs. The difference of them lies in the fusion methods. The result of our best run (hierarchical method _ICTCBCDREL) is shown as Fig 5. By comparing the ICTCBCDREL, ICTCBCDTOA and ICTCBCDAll, we can find that hierarchical method can improve the detection precision obviously. And the integration of 4 modules is very necessary.

References

[1] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, and F. Stentiford, Video copy detection: a comparative study, in CIVR, 2007.

[2] K.I.Lin and C. Yang, “The ANN-Tree: An Index for Efficient Approximate Nearest-Neighbour Search”, In Conf. on Database Systems for Advanced Applications, 2001.

[3] Harris C, Stephens M. A Combined Corner and Edge Detector. Proceedings of Fourth Alvey Vision Conference. 1988, 147-151.

[4] Canny J. A computational approach to edge detection [J].IEEE Transactions on Patten Analysis and Machine Intelligence, 1986, PAMI-8(6):679-698.

[5] Bao Paul, Zhang Lei, Wu Xiaolin. Canny Edge Detection Enhancement by Scale Multiplication [J].IEEE transactions on pattern analysis and machine intelligence, 2005, 27(9):1485-1490.

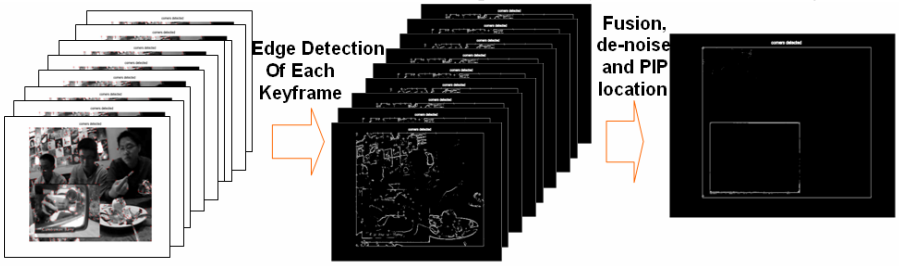

Figure 4: Illustration of PIP boundary location.

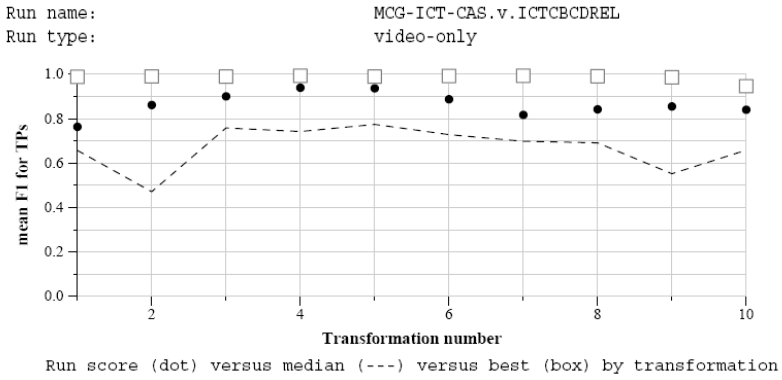

Figure 5: Illustration of our best result comparing with others.

TRECVID 2008 Event Detection By MCG-ICT-CAS*

Junbo Guo, Anan Liu, Yan Song, Zhineng Chen, Lin Pang, Hongtao Xie, Leigang Zhang

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China

{guojunbo, liuanan, songyan, chenzhineng, panglin, xiehongtao, zhangleigang}@ict.ac.cn

ABSTRACT

As for Event Detection in TRECVID 2008, we develop a surveillance system with two parts, the trajectory-based sub-system and the domain knowledge-based sub-system. The former focuses on the research of general methods for event discovery. Human detection and tracking are utilized to generate the trajectory and then novel three-level trajectory features are proposed to detect PersonRuns, PeopleMeet, PeopleSpiltUp, and Embrace. The latter focuses on the study of specified models to improve the results. Based on domain knowledge, three models are respectively constructed for PeopleMeet, Opposingflow, and ElevatorNoEntry. The results are separately shown in the submitted results, "MCG-ICT-CAS_2008_retroED_EVAL08_ENG_s-camera_p-baseline_1" and "MCG-ICT-CAS_2008_retroED_EVAL08_ENG_s-camera_p-Run2_1".

Keywords

Event detection, Surveillance, Trajectory, Domain Knowledge, Human Detection, Tracking.

1. Introduction

Event detection in video surveillance is very important for some public environments (e.g. communities, airport and shopping centers, etc.). However, large amount of video data in surveillance makes it an exhausting work for people to keep watching and finding abnormal events. Therefore, automatic event detection is urgently needed to make the objective, reliable and repeatable decision.

Event detection has been an active research field in recent years. There are mainly two kinds of methods. One is the fundamental method [1] consisting of human detection, tracking and behavior understanding. The current research on the three key problems [2-4] is usually separated and condition-constrained. Therefore, the algorithms are difficult to be implemented in the practical application ideally. The other constructs specific model for the event with spatio-temporal features and detects the event in the video volume [2]. Although machine learning methods [5-6] are widely used to improve the generalization, it is difficult to get a perfect model because of the diversity of patterns.

Since the surveillance video for this task is captured from airport. It is unconstrained and has the characteristics such as highly clutter, massive population flow, heavy occlusion and so on, we find that typical machine learning methods are unsuitable in this situation. As for this practical application, we develop a video surveillance system with two parts, the trajectory-based sub-system and the domain knowledge- based sub-system. The first one implements human detection and tracking to generate trajectory and three-level trajectory features are used to detect PersonRuns, PeopleMeet, PeopleSpiltUp and Embrace. The second one constructs specific models for PeopleMeet, Opposingflow, and ElevatorNoEntry depending on domain knowledge. Therefore, in our exploration for Event Detction in TRECVid 2008, we focus on both generality and specificity to develop a prototype system for video surveillance.

The remainder of the paper is organized as follows. We specifically present the trajectory-based sub-system and the domain knowledge- based sub-system in Section 2 and 3. The experimental results are presented in Section 4.

2. Trajectory-based Sub-system

In this section, we illustrate the trajectory-based sub-system for event detection in video surveillance in details.

2.1 Preprocessing

For each video, we only extracted I and P frame considering both the redundancy in temporal domain and low computational cost. The background subtraction algorithm and morphological operations followed by consist the preprocessing step, only keep those regions that contain more than 30 pixels.

2.2 Human-Detection

The cascade boosting object detection framework in [7] is used for human detection. Specifically, we independently

* This work was supported by National Basic Research Program of China (973 Program, 2007CB311100), and National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), and National Nature Science Foundation of China (60873165, 60802028, 6073056).

train two detectors, full-body and head-shoulder detectors using standard haar-like features. The detection result is made by joint decision of both two detectors. The training data is set as follows. For full-body detector, positive samples are public training data released by DCU, where 3749 people are labeled from 815 images. For head-shoulder detector, positive samples are 3000 frames manually labeled by our team, where 3140 head-shoulders, including frontal, rear and side views, are annotated. The negative samples for both two detectors are manually labeled by our team, which consists of 273 frames without human, collected from the training corpus.

2.3 Human-Tracking

We have tried several state-of-the-art tracking algorithms. Since occlusions happen frequently in limited camera scope, Particle filtering [8] achieves the best performance. Unfortunately, particle filtering is a time-consuming process, especially when the object tracked is large. It is difficult to complete the test on evaluation data within the limited time. Therefore, we adopt the data correlation method with the visual features of the center and color histogram of the detected bounding box.

2.4 Event Detection

It is known that various features can be directly extracted from the trajectory. Then, we proposed a three-level trajectory features for event discovery. From bottom to top, they are individual feature, two-person feature, and crowd feature, as depicted in Fig. 1.

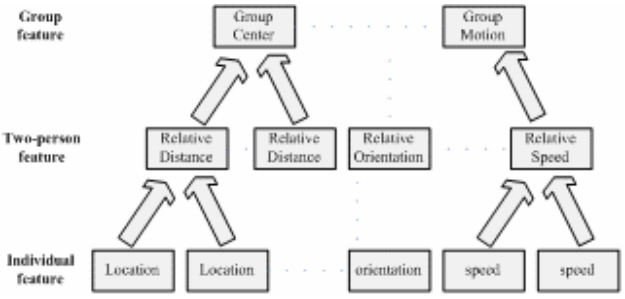

Fig. 1. The three-level hierarchical feature architecture

The individual features are speed and orientation directly extracted from one trajectory. The two-person features are relative speed, relative distance, relative orientation calculated between two persons in the same frame. Crowd features are combinations of two-person features, e.g. the center point of a group of people.

With these features and trajectory information, we set different rules to detect events, including PersonRuns, PeopleMeet, PeopleSpiltUp, and Embrace, as follows:

1) **PersonRuns**: Three types of speed, namely speed between people in two, three and four consecutive frames are extracted from each trajectory. We set three experienced thresholds respectively. Each speed exceeding its corresponding threshold is considered as an available speed. The number of available speed, as well as the corresponding trajectory points is recorded. The decision is made by jointly considering the trajectory length, location of trajectory points and percent of available speed.

2) **PeopleMeet**. PeopleMeet is detected using rules: a): Calculate relative distances between any two individuals in each frame. If there exist a relative distance smaller than a given threshold "D_t", go to step b), otherwise, go to the next frame; b): If two persons appear in the following frames and satisfying rules: their distance decreases continuously, meanwhile, their relative orientation and speed are in reasonable range, which is represented as certain predefined constraints. We decide that the PeopleMeet occures. We go to step a) if relative distance of persons exceeds "D_t" in the above process. Since multi-person meet can be decomposed as some two-person meets, its start time and end time can be determined accordingly.

3) **PeopleSplitUp**. As PeopleSplitUp happens when one or more person separate from a group, our method consists of the following four steps: a): Detect the number of crowds in a frame, using a distance threshold "D_g"; b): Compute and update each crowd center in consecutive frames, recorde the number of frame that each person belongs to a specific crowd, which is called living-time; c): A person is decided to leave the corresponding crowd if the relative distance between him (her) and the crowd center is larger than "D_g", and the living-time of the person is longer than a time threshold "T_g"; d): Track every person belonging to the seperated crowd and PeopleSplitUp event is detected only if there is at least one person coming off the frame.

4) **Embrace**. According to our observation, a large portion of Embrace events happen immediately after PeopleMeet. Therefore we use the trajectory location, relative distance and speed to detect Embrace as follows: Calculating relative distance and speed after PeopleMeet. Embrace is detected when relative distance and speed are respectively below given thresholds. Meanwhile, the trajectory locations of meet persons are nearly unchanged.

3 Domain Knowledge- based Sub-system

In the domain knowledge-based sub-system, we construct three specific methods for ElevatorNoEntry, OpposingFlow and PeopleMeet respectively.

1) **ElevatorNoEntry**. It is obvious that ElevatorNoEntry is related with both the state of the elevator door and the appearance of human. Therefore, we design one detector for the period of door open and close and another for human appearance. Because the elevator doors correspond to fixed regions in the frames and some specific regions change significantly during the period of door open and close, both periods can be detected with the changes of foreground in the door regions. Dynamic background construction and foreground segmentation [9] are adopted here to detect both periods. As for the period between door open and close, we implements human detection and tracking for door region

simultaneously. If the person exists during this period, the event of ElevatorNoEntry occurs.

2) **OpposingFlow**. We detect OpposingFlow as follows: a): Optical flow is calculated depending on Lucas-Kanade algorithm in [10] on a set of densely detected Harris corners, which is derived as low-level features with the post-processing of Guassian smoothing and de-noising; b): Orientation histogram of optical flow in the door region is calculated to represent the statistical feature of optical flow amplitude of corner points. If the value in the bin of reverse direction is over the pre-setted threshold, we mark this frame as a candidate frame; c): To avoid false detection, human detection is implemented in the candidate frame and human tracking is used foreward and backward. d): The candidate is decided to be positive only if the person in current region can be tracked back to last N frames and the trajectory spans over the inside and outside of the door.

3) **PeopleMeet**. Note that both the camera is fixed and the probability of PeopleMeet is varying with regions, we improve PeopleMeet detections by adding a post-process step to trajectory-based sub-system results that gives more weight to some regions containing more people activities when calculate the detectionscore.

Table 1 .Results of Baseline

Events	Ref	Act.P miss	Act.RFA	Recall (%)	Precisi on(%)	Act.D CR	Min.D CR
PersonRuns	314	0.9268	12.5043	7.32	3.474	0.9893	0.9724
PeopleMeet	1182	0.5000	239.5783	50.00	4.605	1.6979	1.0067
PeopleSpilt Up	671	0.5142	178.6417	48.58	3.4479	1.4074	0.9981
Embrace	401	0.8279	84.7907	17.21	1.567	1.2519	0.9993

Table 2.Results of Run2

Events	Ref	Act.P miss	Act.RFA	Recall (%)	Precis ion	Act.D CR	Min .D CR
ElevatorNo Entry	0	NA	0.1174	NA	0	NA	NA
OpposingFl ow	12	0.4167	2.8962	58.33	4.516	0.4311	0.4307
PeopleMeet	1182	0.5964	180.8725	40.36	4.907	1.5008	1.0094

4 Experimental Results

The results of trajectory-based sub-system are considered as the baseline shown in Table 1 and domain knowledge-based sub-system are regarded as Run2 shown in Table 2.

From Table 1 we can see that recall is acceptable and precision is a little low in baseline. The reasons maybe lie in two aspects: 1) The surveillance video is in unconstrained condition and therefore the trajectory features can not perfectly represent the events; 2) The accumulation of errors in human detection, tracking and event detection can have great influence on the final decision. Besides, it is seen that the recalls of PeopleMeet and PeopleSpiltUp is higher than those of PersonRuns and Embrace. It is because the definitions of the former two are more clear and the rules are more robust.

From Table 2, we can see our specific model works well for OppositingFlow. Although the precision is low, the recall and DCR show that our method is effective. The ElevatorNoEntry result is difficult to analyze since there is no reference event annotation, however, our dryrun result as well as our test results on development corpus show that our model is effective. As for PeopleMeet, it is natural that the recall gets lower than that in trajectory-based sub-system, however, the precision only increased from 4.6% to 4.9%, which is lower than our expectation. The possible reason is that giving more weight to the regions containing more people activities also increase the probability of discarding the true detections in other places. To solution this problem, more complicated domain knowledge based rules is necessary.

5 REFERENCES

[1] I. Haritaoglu, D. Harwood, and L. S. Davis, “W: Real-time surveillance of people and their activities,” IEEE Trans. Pattern Anal. Machine Intell., vol. 22, pp. 809–830, Aug. 2000.

[2] Actions as Space-Time Shapes. Blank M., Gorelick L., Shechtman E., Irani M., Basri, R. Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV’05)Volume 2, 17-21 Oct. 2005 Page(s):1395 –1402.

[3] Human detection based on a probabilistic assembly of robust part detectors. K Mikolajczyk, C Schmid, A Zisserman. In Proc. of ECCV, volume 1, pages 69–82, 2004.

[4] Detection and Tracking of Humans by Probabilistic Body Part Assembly. A Micilotta, E Ong, R Bowden. British Machine Vision Conference, Oxford, UK, Sep 2005.

[5] A HMM based semantic analysis framework for sports game event detection. Gu Xu Yu-Fei Ma Hong-Jiang Zhang Shiqiang Yang Proceedings of the IEEE ICIP 2003 Volume: 1, On page(s): 25-28

[6] S Park, JK Aggarwal. A hierarchical Bayesian network for event recognition of human actions and interactions. Multimedia Systems, vol.10, issue.2, papes: 164--179 2004.

[7] Voila P, Jones M. Rapid object detection using adaboosted cascade of simple features. In Proc of IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, USA, pages:511-518, 2001

[8] Arulampalam M S, Maskell S, Gordon N, Clapp T. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Transactions on Signal Processing, vol.50, issue.2, pages: 174-188, 2002.

[9] A.Monnet, A. Mittal, N. Paragios, Visvanathan R. Background modeling and subtraction of dynamic scenes. In the proceeding of ICCV, pages: 1305-1312 vol.2, 2003.

[10] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI*, pages 674–679, 1981.