Zhejiang University at TRECVID 2006

Yanan Liu, Fei Wu, Yueting Zhuang, Shengyi Zhou
Digital media Computing & Design Lab (www.dcd.zju.edu.cn),
College of Computer Science and Technology, Zhejiang University
Zhejiang, Hangzhou, 310027, P.R.China

Abstract. We participated in the high-level feature extraction and interactive-search task for TRECVID 2006. Interaction and integration of multi-modality media types such as visual, audio and textual data in video are the essence of video content analysis. Although any uni-modality type partially expresses limited semantics less or more, video semantics are fully manifested only by interaction and integration of any unimodal. For the high-level feature extraction and interactive-search tasks, taking the temporal-sequenced associated cooccurence characteristic of multimodal media data in video into consideration, we develop a new approach to represent the relations between seperate shots, which mainly uses SimFusion and Locality Preserving Projections (LPP). SimFusion is an effective algorithm to reinforce or propagate the similarity relations between multi-modalities. LPP is an optimal combination of linear and nonlinear dimensionality reduction method. For high-level feature extraction task, e.g.semantic concept detection which actually is a pattern recognition problem, we use SVM as the powerful classifier to perform detection. For interactive search, we make use of relevance feedback to revise search results. We submitted one fun for each task.

1. Introduction

Research in content-based multimedia retrieval is motivated by a growing amount of digital multimedia content in which video data has a big part. Video data comprises plentiful semantics, such as people, scene, object, event and story, etc. Much research effort has been made to negotiate the “semantic gap” between low-level features and high-level concepts. In general, three modalities exist in video, namely the image, audio, and text modalities. How to utilize multi-modality features of video data effectively to better understand the multimedia content remains a great challenge.

A multimodal analysis method for semantic understanding of video includes a fusion step to combine the results of several single media analysis. The two main strategies of fusion are early fusion and late fusion[1]. And most existing methods for video concept detection and video retrieval are based on these two schemes.

As described in [1], early fusion is a scheme that integrates unimodal features before learning concepts, whereas late fusion is a scheme that first reduces unimodal features to obtain separately learned concept scores, then these scores are integrated to learn concepts.

When taking early fusion scheme, unimodal features first extracted. After analysis of the various unimodal streams, the extracted features are combined into a single representation, where simply uses concatenation of unimodal feature vectors to obtain a fused multimedia representation. Early fusion yields a truly multimedia feature representation, but it is still a great difficulty to combine features into a common representation properly and effectively.

In contrast to early fusion, approaches for late fusion learn semantic concepts directly from unimodal features, then combine learned unimodal scores into a multimodal representation. Though late fusion focuses on the individual strength of modalities, the expensiveness in terms of the learning effort of separate supervised learning stage for every modal and an additional learning stage for combination is a big disadvantage.

However, the multimodal media types such as image, audio and text in video are in essence of temporal-sequenced associated occurrence. For instance, during a period of time, although the multi-modality data of continuous video frames, transcripts and audio signal may not occur at once, i.e. asynchronously, they convey the uniform semantics. That is, the multi-modality features extracted from video data present a temporal-sequenced associated cooccurence (TSAC) characteristic, which neither traditional early fusion nor late fusion strategy takes into account.

Several major aspects claim attention when considering TSAC characteristic of video. First, how to propagate similarity correlations between distinct modalities. That is, for some semantics, suppose that a video object presents more similar in one modality, then we need find a way to “re-inforce” the similarities in other modalities based on the given “stronger” similarity. And it is worth notice that the relationships in uni-modality and between multi-modalities are complementary. And the intra-modality similarity can reinforce the inter-modality relationship. Thereby how to effectively propagate corresponding correlations between multi-modalitis is a noticable problem. Secondly, “the curse of dimensionality” has been a well-known problem caused by high dimensionality, which video features inevitably face especially when multi-modalities fuse together. So it is important to find a better dimensionality reduction method. Furthermore, statistical learning methods will be a powerful tool for constructing models.

[2] presents a unified similarity-calculation algorithm SimFusion. This approach uses a Unified Relationship Matrix (*URM*) to represent a set of heterogeneous data objects and their interrelationships. By iteratively computing over the *URM*, SimFusion can effectively integrate relationships from heterogeneous sources when measuring the similarity of two data objects. A Unified Similarity Matrix (*USM*) is defined in this process to represent the similarity values of any data object pairs from same or different data spaces. Thus through SimFusion, we can achieve better results of multi-modality subspace correlation propagation.

As we know, the curse of the dimensionality [6] refers to the fact that in the absence of simplifying assumptions, the sample size needed to estimate a function of several variables to a given degree of accuracy (i.e. to get a reasonably low-variance estimate) grows exponentially with the number of variables.

The problem of dimensionality reduction is introduced as a way to overcome the "curse of the dimensionality" when dealing with vector data in high-dimensional spaces and as a modeling tool for such data [7]. It is defined as the search for a low-dimensional manifold that embeds the high-dimensional data.

Now several techniques for dimensionality reduction have been proposed, usually divide into two parts – linear and nonlinear methods. Linear methods reduce dimension through the use of linear combinations of variable, and nonlinear methods do so with nonlinear functions of variable. The linear combinations can be considered as linear projection, and guaranteed to discover the true structure of data lying on or near a linear subspace of the high-dimensional input space. Principle component analysis (PCA)[8] and projection pursuit[9] are typical methods of this type. Although linear methods are simple to implement, explainable, efficient computable and more extensible, many data sets contain essential nonlinear structure that are invisible to PCA and other linear ways, e.g. the classical "Swiss roll" data set, which intrinsically distribute in a nonlinear manifold. As the research for manifold learning, several traditional non-linear methods have been proposed, such as locally linear embedding (LLE)[10], Isomap[11], and Laplacian eigenmap[12]. All of these algorithms are able to discover the intrinsic nonlinear structure, but they are not able to extend to out-of-sample data directly. That is, they are defined only on the training dataset and it is difficult to evaluate the map for new sample. But then, locality preserving projections (LPP) is a combination of linear and nonlinear aspects.

LPP builds a graph incorporating neighborhood information of the data set. Then using the notion of the Laplacian of the graph, a transformation matrix that maps the data points to a subspace is computed. This linear transformation optimally preserves local neighborhood information in a certain sense. The representation map generated by the algorithm may be viewed as a linear

discrete approximation to a continuous map that naturally arises from the geometry of the manifold. In deed, LPP may be simply applied to any new data point to locate it in the reduced representation space.

In this paper, we propose a new approach for semantic concept detection and topic search in video. Obviously, multi-modality fusion is adopted instead of uni-modality method. For text features, we use Latent Semantic Analysis (LSA) [3] to discover the intrinsic structure of document space. Considering the important temporal-sequenced associated cooccurence characteristic of video, we use SimFusion to propagate correlation from one modality to another, and for much more precise correlations between different modalities. Locality Preserving Projection (LPP) [4] is a noval linear dimensionality reduction algorithm that also shares many of the data representation properties of nonlinear techniques. That is to say, LPP may be simply applied to any new data point to map it in the manifold subapce rather than only defined on the training data set. So we adopt LPP to reduce the high dimension of fused multi-modalities. At last Support Vector Machine (SVM) is used to detect video semantics for high-level feature extraction task; and also we take relevance feedback into consideration for interactive search task.

2. Basic Model Structure

In our approach, a single shot is taken as a basic unit of video semantic concept detection and the interactive search. We perceive of semantic concept detection in video as a pattern recognition problem. Given pattern p, part of shot i, the aim is to obtain a measure, which indicates whether semantic concept w is present in shot i. In the search task, what we aim is to find the closest shots to the query example q and return them to users.

2.1 Low-level Feature Extraction

Low-level features are extracted for each shot. Low-level means the features directly extracted from the source – videos, which distinguish from the high-level semantic concept of video. And the motivation of this paper is to use the labeled training video to classify unkown video into different semantic classes. As video carries multi-modality information including visual, audio, and textual data, the low-level features also compose of three parts.

Image features. A shot is the basic unit; therefore, one key frame within each shot is obtained as a representative image for that shot. Image features are then based on the features extracted from the representative image. There are three different types of image features: color histograms, textures and edges.

Audio features. For each shot, we extract the according audio signal as a "audio clip", and divide the audio clip into overlapped "short-time audio frame". Then a frame feature vector is formed based on the audio features extracted from each audio frame. Because of the variable lengths of shots, we calculate the statistic (mean or variance) of audio frame feature vectors for each shot.

Text features. The source text is the ASR transcript. The dimension of text features is much larger than the other modality features, and text contains abundance of semantic information, therefore we use Latent Semantic Analysis (LSA) to reduce the text dimension and discover the semantic structure. This pre-processing step also reduces the dimension of text features effectively first.

2.2 Multi-modal Subspace Correlation Propagation

As mentioned before, shot is the basic processing unit, so our final result we want is the semantic relationships among shots. However, a shot composes of image, audio and text the three multiple modalities; it is difficult to calculate the similarities among shots directly. Also, the temporal-sequenced associated cooccurence characteristic of video reminds of utilizing the multi-modality relationship propagation to gain much more precise and stable similarities among different shots.

The similarity in same modality is easy to calculated, such as the Euclidean distance between image and image, but the correlation between different modalities is hard to obtain, i.e. the relationship of image and text. Thus SimFusion is an effective way to combine relationships from multiple modalities and achieve multi-modal subspace correlation propagation.

Suppose we have N shots in the training data set X in R^n.

The Unified Relationship Matrix (*URM*) L_{urm} is defined as below:

$$L_{urm} = \begin{vmatrix} l_{11}L_{image} & l_{12}L_{i-a} & l_{13}L_{i-t} & l_{14}L_{i-s} \\ l_{21}L_{a-i} & l_{22}L_{audio} & l_{23}L_{a-t} & l_{24}L_{a-s} \\ l_{31}L_{t-i} & l_{32}L_{t-a} & l_{33}L_{text} & l_{34}L_{t-s} \\ l_{41}L_{s-i} & l_{42}L_{s-a} & l_{43}L_{s-t} & l_{44}L_{shot} \end{vmatrix} . \quad (1)$$

Here L_{image}, L_{audio}, L_{text} and L_{shot} are the intra-modality similarity matrix of image, audio and text spaces respectively. And $L_{i\text{-}a}$, $L_{i\text{-}t}$, $L_{i\text{-}s}$ represent the correlations between image and audio, image and text, image and shot, respectively. The same are the other submatrices. Each submatrix L is $N \times N$. The set of parameters λs are defined to adjust the relative importance of different inter- and intra-modality relationships, and $\sum_{\forall j} l_{ij} = 1, \ \forall i, j, l_{ij} > 0$.

L_{image} and L_{audio} can be calculated based on Euclidean distance, while L_{text} is from Cosine similarity.

Also, the Unified Similarity Matrix (*USM*) is defined as follows:

$$S_{usm} = \begin{vmatrix} 1 & s_{12} & \mathbf{L} & s_{1T} \\ s_{21} & 1 & \mathbf{L} & s_{2T} \\ \mathrm{M} & \mathrm{M} & \mathbf{O} & \mathrm{M} \\ s_{1T} & s_{2T} & \mathbf{L} & 1 \end{vmatrix} \quad . \tag{2}$$

where each element $s_{a,b}$ represents the similarity value between data objects a and b (in this case, between image, audio, text and shot) in the unified space. T is the total number of objects in the unified space, i.e. $T = 4 * N$. It is worth mentioning that the order of data objects presented in S_{usm} and L_{urm} are similar. Having *URM* and *USM* defined, the similarity reinforcement assumption can be represented as :

$$S_{usm}^{new} = L_{urm} S_{usm}^{original} L_{urm}^{T} \quad . \tag{3}$$

Equation (3) is the basic similarity reinforcement calculation in the SimiFusion algorithm. And it can be continued in an iterative manner until the calculation converges or a satisfactory result is obtained, as shown in Equation (4):

$$S_{usm}^{n} = L_{urm} S_{usm}^{n-1} L_{urm}^{T} = L_{urm}^{n} S_{usm}^{0} (L_{urm}^{T})^{n} \quad . \tag{4}$$

In practice, the initial *USM* is often set to be an identity matrix.

The final iterative result S_{usm} can be separated into 4*4 submatrices as L_{urm}. And the last submutrix $W_{N\times N}$ represents the similarity between shots, which is ultimately what we want in this step and will be one input of the next dimension reduction process.

2.3 Dimension Reduction

As mentioned in section 1, LPP is an efficient mean that combines linear and non-linear features of manifold learning.

Given the training data set $X = \{x_1, x_2, \ldots x_N\}$ in R^n as section 2.2, the calculation of LPP will find a transformation matrix A that maps these N points to a set of points $y_1, y_2, \ldots, y_N$ in R^l ($l << n$), such that y_i "represents" x_i, where $y_i = A^T x_i$.

The main procedure of LPP is formally stated below:

1. Choose the weight of the adjacency graph that constructed with the training data set. Here we use $W_{N\times N}$ computed above from SimFusion.

2. Compute the eigenvectores and eigenvalues for the generalized eigenvector problem:

$$XLX^T a = \lambda XDX^T a \; . \tag{5}$$

where D is a diagonal matrix whose entries are column sums of W, $D_{ii} = \sum_j W_{ji}$. $L = D - W$ is the Laplacian matrix. The i^{th} column of matrix X is $\mathbf{x}_i$.

Let the column vectors $\mathbf{a}_0,\ldots,\ \mathbf{a}_{l-1}$ be the solutions of equation (5), ordered according to their eigenvalues, $\lambda_0< \cdots <\lambda_{l-1}$. Thus, the embedding is as follows:

$$\mathbf{x}_i \rightarrow \mathbf{y}_i = A^T\mathbf{x}_i,\ A = (\mathbf{a}_0,\mathbf{a}_1\ldots,\ \mathbf{a}_{l-1})\ . \qquad (6)$$

where $\mathbf{y}_i$ is a l-dimensional vector, and A is a $n\times l$ matrix.

3. High-level Feature Extraction

For high-level feature extraction task, that is semantic concept detection in video, the next step is to train model, which is used as a classifier to detect whether a new shot contains the concept.

Among the large variety of supervised machine learning approaches available, the Support Vector Machine (SVM) framework [16] has proven to be a solid choice. The SVM is able to learn from few examples, handle unbalanced data, and handle unknown or erroneous detected data. An SVM tries to find an optimal separating hyperplane between two classes by maximizing the margin between those two different classes. Finding this optimal hyperplane is viewed as the solution of a quadratic programming problem.

In our approach, we use SVM to construct the classification model. The input takes the features that are processed through above steps in stead of the original data.

4. Interactive Search

For interactive search task, from the previous steps, we have got the similarity relation matrix $W_{N\times N}$ and the LPP coordinates $\mathbf{y}_i$ in the semantic space for shots. Thus for the queries as an example in the training set, we can compare the similarities between the coordinates in the LPP semantic space directly to find the closest results. But for the queries out of the training set, like the processing in high-level feature extraction, we compute their final LPP coordinates in semantic space through the previous steps first, and then compare the new coordinate of query example with other shots.

Indeed, the first results return only by one computation is not the best. We make use of relevance feedback[17] to revise the results from the initial L_{urm}, which is computed through initial S_{usm}. The main purpose of relevance feedback is to revise LPP semantic space and to revise the coordinates of queries out of training set.

In the LPP semantic space, the distance between shots a and b is defined as $d(a,b)$.After one feedback, the set of all the positive examples is represented as

P, and *N* is for the set of all the negative examples. Then, for $\forall a,b \in P$,

$$d(a,b) = a \times d(a,b), \tag{7}$$

where $a < 1$ and is a constant. Meantime, for $\forall a \in P, \forall b \in N$,

$$d(a,b) = b \times d(a,b), \tag{8}$$

where $b > 1$ and is also a constant. These two steps can reduce the distances between positive examples, and enlarge the distances between positive and negative examples based on the information from relevance feedback. Thus this process will revise the LPP semantic space gradually.

For the query example q that is out of training set, suppose there are K positive examples after one feedback, whose LPP coordinates are m_{p1}, m_{p2}, …, m_{pk}. Then the revised coordinate of q is:

$$m_q = \frac{1}{K}(m_{p1} + m_{p1} + \mathbf{L} + m_{pk}) \tag{9}$$

The reason is that as for the K shots are chosen by the user as positive examples, so they are closest to query example with which the uesr is satisfactory. Thus the average of these K positive examples' coordinates is recognized as the coordinate of the query example correctly to some extent.

5. Results and Discussion

For each task, we submitted one run, respectively. It isnecessary to note that it's the first time we participate in Trecvid evaluations, and we do not have much experience, so this year we just focus on algorithm and experiments. And our experiments are only based on the English news data, the Chinese and Arab news data are not used. So the results we submitted are actually from part of the whole test data.

For high-level feature extraction task, figure1 illustrates the comparison between the best, mean and our results. Although our results are from only part of the test data, we can see that our results are close to the median level, and some concepts are even higher than the median level, such as concept 26, 36.

For interactive search task, figure 2 points out the difference between our results and the best, top results. Our search results are not as good as the median level, thus we will enhance the understanding of query topic, the process of relevance feedback and continue to work on the improvement of algorithms.

Figure 1. Comparison with the best and median results for high-level feature extraction

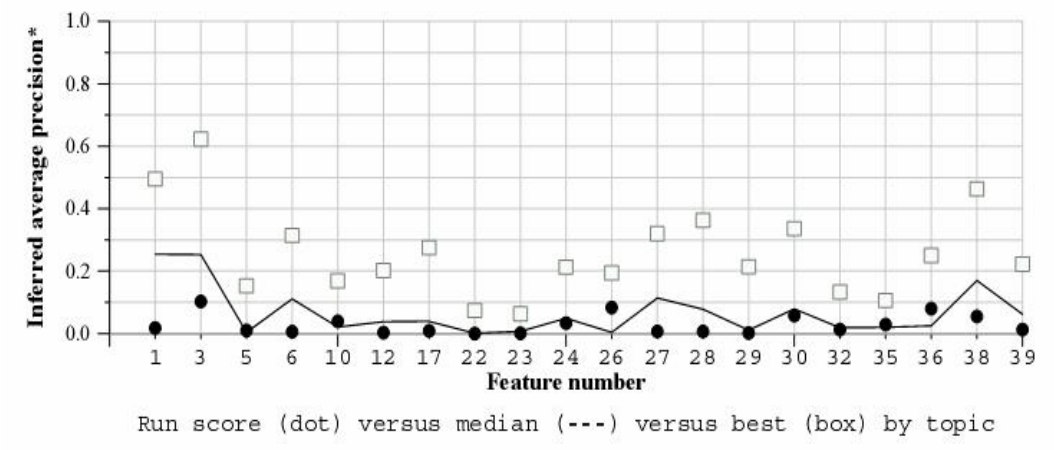
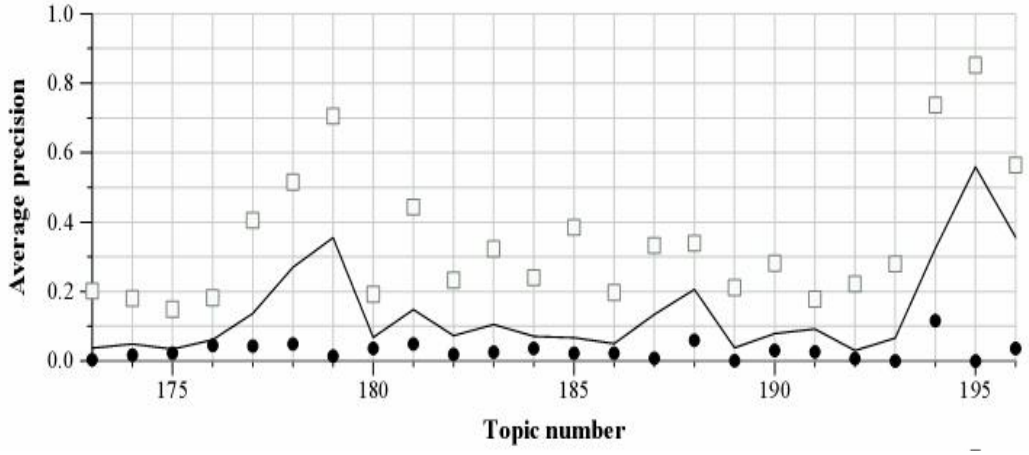


Figure 2. Comparison with the best and median results for interactive search



Semantic understanding of video is a hard but important research topic. The ideas for semantic concept detection and topic retrieval in video are still in process. In this paper, the new approach we present to detect semantic concepts and search topics from video shots is based on SimFusion and LPP. And this method focuses on the temporal-sequenced associated cooccurence characteristic of video. In the future, we plan to design specific processes for different concepts and topics to obtain better results.

References

1. Cees G.M. Snoek, Marcel Worring, Arnold W.M.Smeulders : Early versus Late Fusion in Semantic Video Analysis. Proceedings of the 13th annual ACM International Conference on Multimedia (2005) 399-402
2. Wensi Xi, Edward A.Fox, etc : SimFusion:Measuring Similarity using Unified Relationship Matrix. The 28th Annual International ACM SIGIR Conference (SIGIR'2005)
3. Susan T.Dumais, George W.Furnas, Thomas K.Landauer : Using Latent Semantic Analysis to Improve Access to Textual Information. Proceedings of the SIGCHI conference on Human factors in computing systems (1988) 281-285
4. Xiaofei He, Partha Niyogi : Locality Preserving Projections. Advances in Neural Information Processing Systems (NIPS 2003)
5. Yi Wu, Ching-Yung Lin, Edward Y.Chang, John R.Smith: Multimodal Information Fusion for Video Concept Detection. International Conference on Image Processing (2004) 2391-2394
6. R.Bellman : Adaptive Control Processes: A Guided Tour. Princeton University Press(1961)
7. Miguel Á. Carreira-Perpiñán.: A Review of Dimension Reduction Techniques. Technical report CS-96-09, Dept. of Computer Science, University of Sheffield, UK
8. I.T.Jolliffe : Principal Component Analysis. Springer, New York, 2nd edition (2002)
9. Guy Philip Nason : Design and choice of projection indices. PhD Thesis, University of Bath
10. Sam T. Roweis, Lawrence K. Saul : Nonlinear Dimensionality Reduction by Locally Linear Embedding, Science Vol.290 (2000) 2323-2326
11. Joshua B. Tenenbaum, Vin de Silva, John C. Langford : A Global Geometric Framework for Nonlinear Dimensionality Reduction, Science Vol.290 (2000) 2319-2323
12. Mikhail Belkin, Partha Niyogi : Laplacian Eigenmaps for dimensionality reduction and data representation. Neural Computation, Vol 15, Issue 6 (2003) 1373-1396
13. M.Belkin and P.Niyogi : Laplacian Eigenmaps and spectral techniques for embedding and clustering. Advances in Neural Information Processing Systems 14, MIT Press, Cambridge (2002) 585-591
14. A.Hauptmann, M.Y.Chen, M.Christel, C.Huang, etc: Confounded Expectations: Informedia at TRECVID 2004.
15. C.G.M. Snoek, M.Worring, et al : The MediaMill TRECVID 2004 Semantic Video Search Engine. In Proc. TRECVID Workshop,Gaithesburg, USA (2004)
16. Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
17. Gerard Salton, Chris Buckley: Improving Retrieval Performance by Relevance Feedback. Journal of the American Society for Information Science, 41(4), 288-297, 1990