CLUSTERING-BASED ANALYSIS OF SEMANTIC CONCEPT MODELS FOR VIDEO SHOTS

Markus Koskela[1] and Alan F. Smeaton[1,2]

[1]Centre for Digital Video Processing and [2]Adaptive Information Cluster
Dublin City University, Ireland

ABSTRACT

In this paper we present a clustering-based method for representing semantic concepts on multimodal low-level feature spaces and study the evaluation of the goodness of such models with entropy-based methods. As different semantic concepts in video are most accurately represented with different features and modalities, we utilize the relative model-wise confidence values of the feature extraction techniques in weighting them automatically. The method also provides a natural way of measuring the similarity of different concepts in a multimedia lexicon. The experiments of the paper are conducted using the development set of the TRECVID 2005 corpus together with a common annotation for 39 semantic concepts.

1. INTRODUCTION

The predominant approach to producing large-scale semantic concept models for multimedia data is to treat the problem as a generic learning problem in which training data is used to learn models of different concepts over low-level feature distributions. The set of semantic concepts covered by such models generally form part of a larger ontology and are built independently of each other. This approach is scalable which is a requirement as a comprehensive multimedia lexicon needs to have models for hundreds or thousands of concepts. However the definition of which semantic features are to be modeled tends to be done in terms of information science principles and irrespective of the discriminative power of the semantic concepts. This means that the set of concepts in an ontology may be appealing from an ontological perspective but may contain concepts which have little difference in their discriminative power or there may be large 'gaps' in the resulting overall concept space.

For building concept models, one popular approach is to use discriminative approaches such as support vector machines to classify between positive and negative examples of a certain concept [1]. An alternative is to take a generative approach in which the probability density function of a semantic concept is estimated based on existing training data. In this paper, we follow the latter approach and use global low-level features extracted from the video data, audio track, and keyframe for video shot representation. We study how multimedia concept models built over a clustering method can be interpreted in terms of probability distributions and how the goodness of such models can be assessed with entropy-based methods used in [2]. This approach can also be used for other image or video representations, e.g. latent variable models of local appearance descriptors [3]. The entropy of a certain feature vector's distribution is a measure of how uniformly the used feature distributes the concept over the clusters [4]. We make the assumption that a good model is

such that the distribution is heavily concentrated on only a few clusters, resulting in a low value of entropy. In addition, the similarity of two distributions can be used to measure the overlap of the corresponding concepts. This enables us to produce a similarity matrix for all concepts in an ontology in order to study the inter-concept relations in the lexicon and help us determine the goodness of the overall set of concepts. Inter-concept interaction has been previously studied e.g. in a factor graph framework [5].

The shape of the distribution of a semantic concept over a low-level feature space mapped on a set of clusters depends on factors like the distribution of the original data in the very-high-dimensional pattern space, the feature extraction technique in use, the overall shape of the training set after it has been mapped to the feature space, and the distribution of the studied concept relative to the overall shape of the feature vector distribution. If the feature extraction stage works properly, semantically similar patterns will be mapped in the feature space nearer to each other than semantically dissimilar ones. In the most advantageous situation, the pattern classes might even match clusters in the feature space, i.e. there would exist a one-to-one correspondence between feature vector clusters and pattern classes. With real-world data, this situation is, however, exceedingly rare and the task becomes to measure how well the concept is concentrated in a small cluster subset.

2. CLUSTER ENTROPY AND PERPLEXITY

Given a set of k cluster centroids, we can in theory calculate the a priori probability of each cluster being the best match for any vector $\mathbf{x}$ of the feature space. This is possible if the probability density function (pdf) $p(\mathbf{x})$ is known. The a priori probability of cluster i is

$$P_i = P(\mathbf{x} \in \mathcal{V}_i) = \int_{\mathcal{V}_i} p(\mathbf{x}) \, \mathbf{dx} \ , \qquad (1)$$

where $\mathcal{V}_i$ is its surrounding Voronoi region. With discrete data, we replace the continuous pdf with a discrete probability histogram. Without danger of confusion, the probability can still be denoted as P_i:

$$P_i = P(\mathbf{x} \in \mathcal{V}_i) = \frac{\#\{\, j \mid \mathbf{x}_j \in \mathcal{V}_i \,\}}{N} \ , \qquad (2)$$

where $\#\{\cdot\}$ stands for the cardinality of a set, and N is the size of the training data set whose members are $\mathbf{x}_j$, $j = 0, 1, \ldots, N-1$. Considering a concept $\mathcal{C}_m$ instead of all training data, the probability histogram will be

$$P_i^m = P(\mathbf{x} \in \mathcal{V}_i \mid \mathbf{x} \in \mathcal{C}_m) = \frac{\#\{\, j \mid \mathbf{x}_j \in \mathcal{V}_i, \mathbf{x}_j \in \mathcal{C}_m \,\}}{\#\{\, j \mid \mathbf{x}_j \in \mathcal{C}_m \,\}} \ . \qquad (3)$$

A simple and commonly used measure for the randomness of a symbol distribution is entropy. In our case, the cluster indices for the

vectors of the training set play the role of symbols. The entropy H of a distribution $P = (P_0, P_1, \ldots, P_{k-1})$ is

$$H(P) = -\sum_{i=0}^{k-1} P_i \log P_i \quad , \tag{4}$$

where k is the number symbols in the alphabet of the stochastic information source. P_i is the probability of cluster i being the correct one for an input vector, as defined before. Usually logarithm base of two is used.

If we assume that each of the k clusters is equally probable as the correct one for an input vector, we get the theoretical maximum for the entropy of a clustering $H_{\max^*} = \log k$. In the discrete case, the above definition for the maximum entropy to hold assumes that N is divisible by k. In general this is not the case but the produced error is insignificant with sufficient amount of data, i.e. if $N \gg k$. This can generally be assumed when studying the whole database since the overall aim of clustering is to reduce computational requirements of the retrieval algorithm. With a concept having only a small number of examples available the difference may, however, be considerable so instead of $H_{\max^*}$, we calculate the empirical entropy maximum, $H_{\max}$, for each concept by spreading its distribution over the k clusters as uniformly as possible and using Eq. (4).

Instead of using entropy directly, often a more illustrative measure is perplexity $PPL = 2^H$, commonly utilized e.g. in speech recognition. Perplexity can be considered as the weighted number of equal choices for a random variable; i.e. in this setting, the average number of equivalent clusters that have to be considered given the distribution. Thus, if entropy had its theoretical maximum value $H_{\max^*}$, the perplexity of a clustering would equal the total number of clusters, $PPL_{\max^*} = k$. A suitable performance indicator for feature extraction and the associated clustering methods can be formed by the ratio of perplexity and its maximum value, denoted here as *normalized perplexity*

$$\overline{PPL} = \frac{PPL}{PPL_{\max}} = \frac{2^H}{2^{H_{\max}}} \quad , \tag{5}$$

which is non-negative and ≤ 1 in all cases. In general it can be assumed that clustering distributes the input vectors roughly evenly over the clusters and the normalized perplexity of the whole data should thus be near unity. On the other hand, images with semantic similarity should be mapped to a small cluster subset, provided that the feature extraction and clustering methods have been favorable to that specific concept. In this case, $\overline{PPL}$ should be $\ll 1$.

A straightforward application of $\overline{PPL}$ is use it as a weight of the corresponding distribution in feature fusion. Different multimedia concepts are best represented using multiple features and combining their outputs. A small value of $\overline{PPL}$ corresponds to a well-concentrated distribution, so the relative weight of the corresponding feature should be increased. For example using softmax scaling on the inverse of $\overline{PPL}$, the weight of the ith feature becomes

$$w_i = \frac{\exp(1/\overline{PPL}_i)}{\sum_j \exp(1/\overline{PPL}_j)} \quad . \tag{6}$$

3. INTER-CONCEPT SIMILARITY

When considering the multiple concepts in a lexicon, an interesting question is the similarity between two concepts. Continuing with the information-theoretic approach, a natural measure of two concepts' similarity is their mutual information. Let us denote by P^m

and P^n the probability distributions of concepts $\mathcal{C}_m$ and $\mathcal{C}_n$. As entropy measures the randomness of a distribution, mutual information $I(P^m, P^n)$ can be used for studying the interplay between two distributions

$$I(P^m, P^n) = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} P_{ij}^{mn} \log \frac{P_{ij}^{mn}}{P_i^m P_j^n} \quad , \tag{7}$$

where P^{mn} is the estimated joint probability of the two concepts. Using mutual information as a measure of similarity for different feature extraction methods was examined in [4].

However, when using mutual information in estimating inter-concept similarities, sparse data can be a problem. In order to obtain an accurate enough model of a concept, k has to be relatively large, resulting in a sparse joint probability matrix P^{mn} unless we have a lot of training data. Therefore, we take a different approach and use a bin-to-bin histogram distance measure in estimating the concept similarities. A number of such measures are available, including intersection, Euclidean distance, χ^2-statistic, and Kullback-Leibler divergence. In this paper, we use Jeffrey divergence [6]

$$d_{\mathrm{JD}}(P^m, P^n) = \sum_{i=0}^{k-1} \left(P_i^m \log \frac{P_i^m}{\hat{P}_i} + P_i^n \log \frac{P_i^n}{\hat{P}_i} \right) \quad , \tag{8}$$

where $\hat{P} = (P^m + P^n)/2$ is the mean distribution, as it is symmetric and numerically stable with empirical distributions.

4. EXPERIMENTS

In the following experiments, we use the development set of the TRECVID 2005 [7] corpus consisting of about 80 hours of TV news recorded in November 2004. After automatic shot boundary detection [8], the data set contains 43 907 shots. A joint effort to the TRECVID participants was organized to annotate the whole development set for 39 semantic concepts developed in the ARDA/NRRC workshop on Large Scale Ontology for Multimedia (LSCOM), listed in Table 1. For more detailed descriptions of the concepts and their manual annotation see [7] and [9]. Most of the development set was annotated twice, so we adopted a rule that a shot is considered relevant if either one of the annotations had marked it so.

As low-level features, we used two video features (MPEG-7 Motion Activity (MA) and temporal color moments (CM)), three MPEG-7 image descriptors calculated from the main shot keyframe (Color Layout (CL), Edge Histogram (EH), and Homogeneous Texture (HT)), and one audio feature (Mel-scaled cepstral coefficient (CE)). For more details on these, see [10]. We used the Self-Organizing Map (SOM) [11] as the clustering method with $k = 256$ (16×16 map units) for all features. This was purely for convenience as we used the same clustering runs in [10], and any hard clustering method would be applicable. In fact, since the SOM algorithm is a trade-off between clustering and preserving topology, common clustering algorithms such as k-means often perform slightly better than the SOM when map topology is ignored [2].

4.1. Normalized perplexity

The lowest and highest $\overline{PPL}$ values of the six features for the 39 LSCOM concepts are listed in Table 1. Examining these we can see that on concepts *boat/ship*, *desert*, *maps*, *snow*, and *animal*, distributions of the best features are most concentrated. On the other hand, concepts with the most uniformly distributed best clusterings are *person*, *face*, *outdoor*, *walking/running*, and *building*. Overall,

concept	size	$\overline{PPL}$	feat.	$\overline{PPL}$	feat.
airplane	347	0.20	EH	0.46	MA
animal	380	0.19	CM	0.46	MA
boat/ship	282	0.14	CM	0.38	MA
building	3213	0.64	CM	0.84	CE
bus	110	0.49	CM	0.60	EH
car	2932	0.58	CM	0.82	MA
charts	405	0.23	CM	0.49	CE
computer/tv screen	1890	0.22	EH	0.54	CE
corporate leader	1031	0.47	MA	0.65	CE
court	133	0.26	CL	0.49	MA
crowd	4443	0.48	EH	0.80	CE
desert	270	0.16	EH	0.36	CE
entertainment	5038	0.51	CE	0.84	MA
explosion/fire	447	0.29	CM	0.57	MA
face	27751	0.81	MA	0.95	CE
flag us	305	0.20	EH	0.44	CE
government leader	3447	0.59	EH	0.77	CE
maps	819	0.16	HT	0.38	CE
meeting	1793	0.37	MA	0.60	CE
military	1592	0.42	CM	0.77	MA
mountain	554	0.22	EH	0.52	MA
natural disaster	268	0.22	CM	0.32	MA
office	591	0.41	CM	0.51	CL
outdoor	14463	0.74	CM	0.92	CE
people marching	839	0.26	EH	0.55	CE
person	29440	0.84	MA	0.96	CE
police/security	297	0.32	CM	0.41	EH
prisoner	74	0.41	CL	0.70	MA
road	2817	0.56	CM	0.83	CE
sky	3739	0.45	CM	0.84	CE
snow	138	0.17	CM	0.50	MA
sports	1761	0.41	CM	0.59	HT
studio	5080	0.20	CM	0.46	CE
truck	341	0.33	EH	0.47	CE
urban	3831	0.61	CM	0.86	MA
vegetation	1680	0.51	EH	0.77	MA
walking/running	3701	0.64	MA	0.86	CE
waterscape/waterfr.	1441	0.30	CM	0.73	MA
weather	595	0.22	CE	0.48	MA
all shots	43907	0.90	MA	0.97	CE

Table 1. The lowest and highest $\overline{PPL}$ values for the concepts.

we see that common concepts tend to have higher values of $\overline{PPL}$, which to a certain level is a direct consequence of the larger number of relevant shots which inhabit more clusters. On the extreme, 67% and 63% of the shots in the collection are relevant for concepts *person* and *face*, respectively, making it extremely hard to build effective models for such generic concepts. Finally, the $\overline{PPL}$ results for all shots show that the assumption that the clustering method distributes the data evenly is not completely satisfied, with the Motion Activity feature producing the most nonuniform clustering. Consequently, if the data is very unevenly distributed, it has an effect on the relative confidence values and should be taken into account in determining the feature-wise weights.

In addition, Table 1 lists the features that yield the lowest and highest $\overline{PPL}$ values. It can be seen that each feature yields the lowest $\overline{PPL}$ value for at least one concept, highlighting the need for using diverse features for modeling multimedia concepts and fusing information from the multiple modalities of video data.

4.2. Inter-concept similarity

In the second experiment we study the inter-concept similarities of the 39 LSCOM concepts. We use a linear combination of the six multimodal features, weighted based on Eq. (6). The similarities between concepts in the six clusterings are measured using Jeffrey divergence (Eq. (8)). A full matrix of inter-concept similarities would be difficult to illustrate due to the relatively large number of concepts. Therefore, Table 2 lists instead the five most similar concepts for each of the 39 concepts and a concept dendrogram built using weighted pair-group average linkage.

5. CONCLUSIONS

In this paper we present a method for estimating the goodness of a semantic concept model over a clustering in the low-level feature space. This can be used directly in assessing the reliability of the model and the coverage of the set of semantic concepts in terms of the underlying low-level features. An important aspect of our work is the fact that we deal with distributions over common feature spaces and set of clusters instead of common data items, enabling us to compare concepts trained with different datasets. Extensive annotations over large amounts of multimedia data are rare and laborious to produce, so it is beneficial to be able to use existing annotated datasets to analyze also completely new data. On the other hand, the presented method is readily scalable to large multimedia lexicons as each concept model is represented as a set of distributions over common clusterings in the used feature spaces. Adding a new concept thus requires only the estimation of the distributions on the feature-wise clusterings.

The number of clusters, k, is an important parameter for any clustering-based method and depends on the task at hand. In this application, however, finding an optimal value for k is difficult as an objective evaluation of the similarities between semantic concepts is impossible as everyone has her own subjective views on different concepts. Producing a useful ground-truth would also require the collection of large amount of questionnaires. Overall, coarser representations of concept distributions are useful for concepts for which less training data are available, for initial feature filtering, and for measuring similarities between concepts. A larger value for k is likely to be needed when using cluster distribution models for tasks like automatic annotation or concept detection.

In future work, an important topic is the utilization of the estimated inter-concept relationships in video indexing and retrieval. Concept models can be effectively used as mid-level features in retrieval as they can be trained off-line with considerably more positive and negative examples than what are typically available on-line for an ordinary multimedia query. Furthermore, as the presence of a concept may often reduce the probability of certain other concepts (e.g. *desert* and *snow*), one issue is to study the utilization of the lack of a concept as well as its presence, either as negative models or models of semantic concepts' negatives.

6. REFERENCES

[1] M.R. Naphade and J.R. Smith, "Learning visual models of semantic concepts," in *Proceedings of International Conference on Image Processing (ICIP 2003)*, Barcelona, Spain, September 2003, vol. 2, pp. 531–534.

[2] M. Koskela, J. Laaksonen, and E. Oja, "Entropy-based measures for clustering and SOM topology preservation applied to content-based image indexing and retrieval," in *Proceedings of*

concept	five most similar concepts
face	person, government leader, outdoor, building, walking/running
person	face, outdoor, government leader, walking/running, building
government leader	face, person, meeting, outdoor, building
corporate leader	face, person, government leader, meeting, outdoor
meeting	government leader, face, person, building, crowd
outdoor	urban, building, road, walking/running, car
urban	outdoor, building, road, car, walking/running
building	urban, outdoor, road, car, person
car	road, outdoor, urban, building, walking/running
road	urban, car, outdoor, building, walking/running
crowd	walking/running, outdoor, urban, people marching, person
walking/running	outdoor, urban, crowd, road, person
vegetation	outdoor, building, walking/running, urban, road
military	outdoor, urban, building, walking/running, road
sky	outdoor, building, urban, road, military
entertainment	person, face, outdoor, walking/running, urban
sports	walking/running, outdoor, car, vegetation, person
office	person, face, outdoor, entertainment, building
people marching	crowd, walking/running, outdoor, urban, military
police/security	crowd, walking/running, urban, outdoor, road
natural disaster	building, urban, outdoor, road, military
mountain	sky, waterscape/waterfront, outdoor, road, car
waterscape/waterfront	sky, outdoor, mountain, boat/ship, building
boat/ship	waterscape/waterfront, sky, mountain, road, outdoor
desert	sky, mountain, explosion/fire, waterscape/waterfront, outdoor
explosion/fire	sky, outdoor, urban, military, building
airplane	sky, waterscape/waterfront, outdoor, road, building
truck	road, car, urban, outdoor, building
animal	waterscape/waterfront, outdoor, sky, car, road
snow	mountain, sky, waterscape/waterfront, airplane, animal
computer/tv screen	studio, face, person, meeting, building
studio	computer/tv screen, face, person, meeting, maps
maps	weather, studio, face, person, charts
weather	maps, person, face, charts, outdoor
charts	weather, person, face, maps, computer/tv screen
flag us	government leader, face, person, crowd, walking/running
bus	road, car, urban, building, outdoor
court	meeting, government leader, person, face, corporate leader
prisoner	military, government leader, person, face, walking/running

Table 2. Five most similar concepts for each of the 39 LSCOM concepts and a dendrogram of concept relationships.

17th International Conference on Pattern Recognition (ICPR 2004), Cambridge, UK, August 2004, vol. 2, pp. 1005–1008.

[3] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering object categories in image collections," in *Proceedings of the International Conference on Computer Vision (ICCV 2005)*, Beijing, China, October 2005, pp. 370–377.

[4] J. Laaksonen, M. Koskela, and E. Oja, "Class distributions on SOM surfaces for feature extraction and object retrieval," *Neural Networks*, vol. 17, no. 8-9, pp. 1121–1133, October-November 2004.

[5] M.R. Naphade, I. Kozintsev, and T. Huang, "A factor graph framework for semantic video indexing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 1, pp. 40–52, January 2002.

[6] J. Puzicha, T. Hofmann, and J. Buhmann, "Non-parametric similarity measures for unsupervised texture segmentation and image retrieval," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'97)*, San Juan, Puerto Rico, June 1997, pp. 267–272.

[7] P. Over, T. Ianeva, W. Kraaij, and A.F. Smeaton, "TRECVID 2005 - an introduction," in *TREC Video Retrieval Evaluation Online Proceedings*. TRECVID, 2005.

[8] C. Petersohn, "Fraunhofer HHI at TRECVID 2004: Shot boundary detection system," in *TREC Video Retrieval Evaluation Online Proceedings*. TRECVID, 2004.

[9] T. Volkmer, J.R. Smith, A. Natsev, M. Campbell, and M. Naphade, "A web-based system for collaborative annotation of large image and video collections," in *Proceedings of the 13th ACM International Conference on Multimedia*, Singapore, November 2005, pp. 892–901.

[10] M. Koskela, J. Laaksonen, M. Sjöberg, and H. Muurinen, "PicSOM experiments in TRECVID 2005," in *TREC Video Retrieval Evaluation Online Proceedings*. TRECVID, 2005.

[11] T. Kohonen, *Self-Organizing Maps*, Springer-Verlag, 2001.